Juan D. Velásquez
Sebastián A. Ríos
Robert J. Howlett
Lakhmi C. Jain (Eds.)

LNAI 5712

Knowledge-Based and Intelligent Information and Engineering Systems

13th International Conference, KES 2009
Santiago, Chile, September 2009
Proceedings, Part II

2 Part II

KES

Springer

Lecture Notes in Artificial Intelligence 5712

Edited by R. Goebel, J. Siekmann, and W. Wahlster

Subseries of Lecture Notes in Computer Science

Juan D. Velásquez   Sebastián A. Ríos
Robert J. Howlett   Lakhmi C. Jain (Eds.)

Knowledge-Based and Intelligent Information and Engineering Systems

13th International Conference, KES 2009
Santiago, Chile, September 28-30, 2009
Proceedings, Part II

Springer

Volume Editors

Juan D. Velásquez
Sebastián A. Ríos
University of Chile
Republica 701, Santiago, Chile, 8370439
E-mail: {jvelasqu,srios}@dii.uchile.cl

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

Lakhmi C. Jain
University of South Australia
Mawson Lakes, SA, 5095, Australia
E-mail: Lakhmi.Jain@unisa.edu.au

Library of Congress Control Number: 2009935030

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

LNCS Sublibrary: SL 7 – Artificial Intelligence

ISSN	0302-9743
ISBN-10	3-642-04591-X Springer Berlin Heidelberg New York
ISBN-13	978-3-642-04591-2 Springer Berlin Heidelberg New York

springer.com

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 12766178 06/3180 5 4 3 2 1 0

Preface

On behalf of KES International and the KES 2009 Organising Committee we are very pleased to present these volumes, the proceedings of the 13th International Conference on Knowledge-Based Intelligent Information and Engineering Systems, held at the Faculty of Physical Sciences and Mathematics, University of Chile, in Santiago de Chile.

This year, the broad focus of the KES annual conference was on intelligent applications, emergent intelligent technologies and generic topics relating to the theory, methods, tools and techniques of intelligent systems. This covers a wide range of interests, attracting many high-quality papers, which were subjected to a very rigorous review process. Thus, these volumes contain the best papers, carefully selected from an impressively large number of submissions, on an interesting range of intelligent-systems topics.

For the first time in over a decade of KES events, the annual conference came to South America, to Chile. For many delegates this represented the antipode of their own countries. We recognise the tremendous effort it took for everyone to travel to Chile, and we hope this effort was rewarded. Delegates were presented with the opportunity of sharing their knowledge of high-tech topics on theory and application of intelligent systems and establishing human networks for future work in similar research areas, creating new synergies, and perhaps even, new innovative fields of study. The fact that this occurred in an interesting and beautiful area of the world was an added bonus.

The year 2009 stands out as being the year in which the world's financial situation has impacted on the economies of most countries. This has made it difficult to develop meetings and conferences in many places. However, we are really happy to see the KES conference series continue to be an attractor engine for many researchers, PhD students and scholars in general, despite its location in a very far away country like Chile.

We are grateful to many friends and colleagues for making the KES 2009 conference happen. Unfortunately the list of contributors is so long that it would be difficult to include every single one of them.

However, we would like to express our appreciation of the Millennium Institute of Complex Engineering Systems, whose financial support made possible the local organization, the Millennium Scientific Initiative of the Chilean Government, the Department of Industrial Engineering (DIE), and the Faculty of Physical Sciences and Mathematics of the University of Chile.

Also we would like to acknowledge the work of Victor Rebolledo and Gaston L'Huillier, who were in charge of local organisation, and the team of DIE students and Juan F. Moreno, local general organiser, all of whom worked hard to make the conference a success.

We would like to thank the reviewers, who were essential in providing their reviews of the papers. We are very grateful for this service, without which the conference would not have been possible. We thank the high-profile keynote speakers for providing interesting and informed talks to provoke subsequent discussions.

An important distinction of the KES conferences over others is the Invited Session Programme. Invited Sessions give new and established researchers an opportunity to present a "mini-conference" of their own. By this means they can bring to public view a topic at the leading edge of intelligent systems. This mechanism for feeding new blood into research is very valuable. For this reason we must thank the Invited Session Chairs who contributed in this way.

In some ways, the most important contributors to the conference were the authors, presenters and delegates, without whom the conference could not have taken place. So we thank them for their contribution.

We hope the attendees all found KES 2009 a marvellous and worthwhile experience for learning, teaching, and expanding the research networks and that they enjoyed visiting Chile. We wish that readers of the proceedings will find them a useful archive of the latest results presented at the conference and a source of knowledge and inspiration for their research.

September 2009

Juan D. Velásquez
Sebastián A. Ríos
Robert J. Howlett
Lakhmi C. Jain

Organisation

Conference Committee

General Chairs

Juan D. Velásquez
Department of Industrial Engineering
Faculty of Mathematics and Physical Sciences
University of Chile, Santiago, Chile

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

KES International Executive Chair

Robert J. Howlett
Center for Smart Systems
University of Brighton, UK

KES 2009 Session Chair Organiser

Sebastián A. Rios
Department of Industrial Engineering
Faculty of Mathematics and Physical Sciences
University of Chile, Santiago, Chile

Local Organising Committee

Victor Rebolledo
Sebastián A. Ríos
Juan D. Velásquez
Department of Industrial Engineering
Faculty of Mathematics and Physical Sciences
University of Chile, Santiago, Chile

KES International Operations Manager

Peter J. Cushion

Proceedings Assembling Team

Gaston L'Huillier
Victor Rebolledo
Department of Industrial Engineering
Faculty of Mathematics and Physical Sciences
University of Chile, Santiago, Chile

International Programme Committee

Abe, Akinori	ATR Knowledge Science Laboratories, Japan
Abe, Jair M.	University of Sao Paulo, Brazil
Adachi, Yoshinori	Chubu University, Japan
Angelov, Plamen	University of Lancaster, UK
Arroyo-Figueroa, Gustavo	Instituto de Investigaciones Electricas, Mexico
Baba, Norio	Osaka Kyoiku University, Japan
Balachandran, Bala M.	Camberra University, Australia
Balas, Valentina Emilia	University of Arad, Romania
Bandyopadhyay, Sanghamitra	Indian Statistical Institute, India
Bichindarit, Isabelle	University of Washington, USA
Boicu, Mihai	George Mason University, USA
Bottema, Murk	Flinders University, Australia
Braga, Antonio de Padua	UFMG, Brazil
Brahnam, Sheryl	Missouri State University, USA
Breuel, Thomas	German Research Center for Artificial Intelligence, DFKI GmbH, Germany
Brna, Paul	University of Glasgow, UK
Butz, Cory	University of Regina, Canada
Cali, Andrea	Oxford University, UK
Camastra, Francesco	University of Naples Parthenope, Italy
Castellano, Giovanna	University of Bari, Italy
Chan, Chien-Chung	The University of Akron, USA
Chen, Yen-Wei	Ritsumeikan University, Japan
Cuzzocrea, Alfredo	University of Calabria, Italy
Da Silva, Ivan Nunes	University of São Paulo, Brazil
Dengel, Andreas	German Research Center for Artificial Intelligence (DFKI) GmbH, Germany
Devanathan, R.	Hindustan College of Engineering, India
Elomaa, Tapio	Tampere University of Technology, Finland
Er, Meng Joo	School of Electrical & Electronic Engineering, Malaysia
Fasli, Maria	University of Essex, UK
Felfernig, Alexander	University of Klagenfurt, Austria
Feng, Jun	Hohai University, China
Franco, Leonardo	UK
García-Sebastián, Maite	UPV/EHU, Spain
George, Jieh-Shan	Providence University, Taiwan
Ghosh, Ashish	Indian Statistical Institute, India
Godoy, Daniela	UNICEN, Argentina
Graña, Manuel	Universidad País Vasco, Spain
Ha, Sung Ho	Kyungpook National University, South Korea
Hatzilygeroudis, Ioannis	University of Patras, Greece
Herrera, Francisco	University of Granada, Spain

Hintea, Sorin	Technical University of Cluj-Napoca, Romania
Holmes, Dawn	University of California at Santa Barbara, USA
Honda, Katsuhiro	Osaka Prefecture University, Japan
Hong, Tzung-Pei	National University of Kaohsiung, Taiwan
Horng, Mong-Fong	Nat. Cheng Kung University, Taiwan
Hu, Sanqing	Mayo Clinic College of Medicine, USA
Huang, Guang-Bin	Nanyang Technological University, Singapore
Ines Pena de Carrillo, Clara	Universidad Industrial de Santander, Colombia
Inuiguchi, Masahiro	Osaka University, Japan
Inuzuka, Nobuhiro	Nagoya Institute of Technology, Japan
Ishibuchi, Hisao	Osaka Prefecture University, Japan
Ishida, Yoshiteru	Toyohashi University of Technology, Japan
Ishii, Naohiro	Aichi Institute of Technology, Japan
István, Vassányi	University of Pannonia, Hungary
Ito, Takayuki	Nagoya Institute of Technology, Japan
Iwahori, Yuji	Chubu University, Japan
Jain, Lakhmi	University of South Australia
Jannach, Dietmar	Technische Universität Dortmund, Germany
Kaczmar, Urszula Markowska	Wroclaw University of Technology, Poland
Kanda, Taki	Bunri University of Hospitality, Japan
Kastania, Anastasia	Athens University of Economics and Business, Greece
Klawonn, Frank	University of Applied Sciences Braunschweig, Germany
Koczkodaj, Waldemar W.	Laurentian University, Canada
Kodogiannis, Vassilis	University of Westminster, UK
Koenig, Andreas	Technische Universität Kaiserslautern, Germany
Kojiri, Tomoko	Nagoya University, Japan
Konar, Amit	Jadavpur University, India
Kovalerchuk, Boris	Central Washington University, USA
Kusiak, Andrew	University of Iowa, USA
Kwasnicka, Halina	Wroclaw University of Technology, Poland
Lee, Geuk	Hannam University, South Korea
Lee, Huey-Ming	Chinese Culture University, Taiwan
Lensu, Anssi	University of Jyväskylä, Finland
Lin, Lily	China University of Technology, Taiwan
Liszka, Kathy J.	The University of Akron, USA
Liu, James	The Hong Kong Polytechnic University, China
Loucopoulos, Pericles	Loughborough University, UK
Lovrek, Ignac	University of Zagreb, Croatia
Lygouras, John N.	Democritus University of Thrace, Greece
Markey, Mia	The University of Texas, USA

Mital, Dinesh P.	University of Medicine & Dentistry of New Jersey, USA
Montani, Stefania	Università del Piemonte Orientale, Italy
Mora, Manuel	Autonomous University of Aguascalientes, Mexico
Moraga, Claudio	European Centre for Soft Computing, Spain
Mukai, Naoto	Tokyo University of Science, Japan
Mumford, Christine L.	Cardiff University, UK
Nasraoui, Olfa	University of Louisville, USA
Nauck Detlef D.	BT, UK
Nguyen, Ngoc	Wroclaw University of Technology, Poland
Niskanen, Vesa A.	University of Helsinki, Finland
Ogiela, Lidia	AGH University of Science and Technology, Poland
Ogiela, Marek	AGH University of Science and Technology, Poland
Ohsawa, Yukio	University of Tokyo, Japan
‘O’Hare, Gregory	UCD School of Computer Science and Informatics, Ireland
Palade, Vasile	Oxford University, UK
Park, Kwang-Hyun	KAIST, South Korea
Percannella, Gennaro	Università di Salerno, Italy
Petrosino, Alfredo	University of Naples, Italy
Phillips-Wren, Gloria	Loyola College in Maryland, USA
Pratihar, Dilip Kumar	Indian Institute of Technology, Kharagpur, India
Reidsema, Carl	University of New South Wales, Australia
Remagnino, Paolo	Kingston University, UK
Resconi, Germano	Catholic University in Brescia, Italy
Rhee, Phill Kyu	Inha University, Korea
Rodriguez, Marko A.	Los Alamos National Laboratory, USA
Sampaio, Paolo	University of Madeira, Portugal
Sansone, Carlo	Università degli Studi di Napoli Federico II, Italy
Sato-Ilic, Mika	University of Tsukuba, Japan
Sawicki, Dariusz	Warsaw University of Technology, Poland
Setchi, R.	Cardiff University, UK
Silverman, Barry G.	University of Pennsylvania, USA
Sordo, Margarita	Harvard Medical School, USA
Szczerbicka, Helena	Leibniz University Hanover, Germany
Szczerbicki, Edward	University of Newcastle, Australia
Tanaka, Mieko	Tottori University, Japan
Tanaka, Takushi	Fukuoka Inst. Tech, Japan
Tecuci, Gheorghe	George Mason University, USA
Thalmann, Daniel	EPFL Vrlab, Switzerland
Tolk, Andreas	Old Dominion University, USA

Toro, Carlos Andrés	VICOMTech, Spain
Tsihrintzis, George	University of Pireaus, Greece
Turchetti, Claudio	Università Politecnica delle Marche, Italy
Ushiama,Tatetoshi	Kyushu University, Japan
Vakali, Athena	Aristotle University of Thessaloniki, Greece
Velásquez, Juan D.	University of Chile
Vellido, Alfredo	Universitat Politècnica de Catalunya, Spain
Vialatte, Francois-B.	Riken BSI, Lab. ABSP, Japan
Virvou, Maria	University of Pireaus, Greece
Wang, Guoren	Northeastern University, China
Wang, Justin	LaTrobe University, Australia
Watada, Junzo	Waseda University, Japan
Watanabe, Toyohide	Nagoya University, Japan
Watkins, Jennifer H.	Los Alamos National Laboratory, USA
Weber, Richard	University of Chile
Weber, Rosina	The iSchool at Drexel University, USA
Williams, M. Howard	Heriot-Watt University, UK
Windeatt, Terry	University of Surrey, UK
Xiang, Yang	University of Guelph, Canada
Yoshida, Hiroyuki	Harvard Medical School, USA
Younan, Nick	Mississippi State University, USA
Zazula, Damjan	University of Maribor, Slovenia
Zhang, Minjie	University of Wollongong, Australia
Zhang, Zili	Deakin University, Australia
Zharkova, Valentina	University of Bradford, UK
Zinsmeister, Stefan	University of Kaiserslautern, Germany

General Track Programme Committee

Bruno Apolloni
Bojana Basic
Floriana Esposito
Anne Håkansson
Miroslav Karny
Honghai Liu
Ngoc Nguyen
Andreas Nuernberger
Tuan Pham
Bernd Reusch
Sebastián Ríos

Invited Sessions Programme Committee

Akinore Abe
Yoshinori Adache
Norio Baba
Bala Balachanan
Alfredo Cuzzocrea
Manuel Grana
Anne Håkansson
Sorin Hintea
Yoshiteru Ishida
Yuji Iwahori
Lakhmi Jain
Urszula Kaczmar
Halina Kwasncika
Hsuan-Shih Lee
Mark Liao
Kazumi Nakamatsu
Sebastián Ríos
Kazuhiko Tusuda
Jeffrey Tweedale
Athena Vakali
Juan D. Velásquez
Junzo Watada
Toyohide Watanabe
Katsutoshi Yada

KES 2009 Reviewers

Abdel-Badeeh Salem
Adam Nowak
Adam Slowik
Akihiro Hayashi
Akinori Abe
Alex Hariz
Alexander Felfernig
Alfredo Cuzzocrea
Alfredo Petrosino
Alis Bielan
Anastasia Kastania
Andreas Dengel
Andreas Koenig
Andreas Tolk
Andrzej Sluzek
Angelina Tzacheva
Ani Amizic
Ani Grubisic
Anna Costa
Anne Håkansson
Artemis Hatzigeorgiou
Athena Vakali
Aytul Ercil
Bala Balachandran
Bernd Reusch
Bojana Dalbelo Basic
Boria Vrdoljak
Boris Kovalerchuk
Bosko Bekavac
Bruno Apolloni
Bruno Pouliquen
C.P. Lim
Carl Reidsema
Carlo Sansone
Carsten Saathoff
Chia-Tong Tang
Chien-Chung Chan
Chih-Wen Su
Claudio De Stefano
Claudio Moraga
Colin Fyfe
Cosimo Palmisano
Cuong To
Damir Cavar
Damjan Zazula
Daniela Godoy
Dario Malchiodi
Dariusz Sawicki
Dat Tran
Dau Fuji
David Lee
Davor Grgic
Davor Skrlec
Dawn Holmes
Demetri Terzopoulos
Diana Simic
Dietmar Jannach
Dilip Pratihar
Dingsheng Wan
Don Jeng
Donggang Yu
Dragan Gamberger
Dragan Jevtic
Ebrahim Al-Hashel
Eisuke Itoh
Emir Imamagic
Eugene Hsiao
Federico Pedersini
Felipe Aguilera
Feng-Tse Lin
Francesco Camastra
Francisco Herrera
Franco Pedreschi
Francois Vialatte
Franz Leberl
Fred Nicolls
Gabriel Oltean
George Tsihrintzis
Germano Resconi
Giancarlo Iannizzotto
Giorgio Valentini
Gloria Philips-Wren
Gordan Gledec
Gordan Jezic
Gustavo Arroyo-Figueroa
Halina Kwasnicka
Hans Jørgen Andersen
Haruki Kawanaka
Hideaki Ito
Hideo Funaoi
Hiro Yoshida
Hiroyuki Mitsuhara
Hisao Ishibuchi
Hisatoshi Mochizuki
Honghai Liu
Hsuan-Shih Lee
Huerta Huerta
Huey-Ming Lee
Ida Raffaelli
Igor Mekterovic
Ioannis Hatzilygeroudis
Isabelle Bichindaritz
Ivan Silva
Ivan Villaverde
Jair Abe
James Liu
James Peters
Jan Snajder
Janez Demsar
Jason Wang
Jeffrey Tweedale
Jennifer Watkins
Jeremiah Deng
Jing-Long Wang
John Fader
John Hefferan
John Lygouras
Josipa Kern
Juan D. Velásquez
Juliusz Kulikowski
Jun Feng
Junzo Watada
Karolj Skala
Katarina Curko
Kathy Liszka
Katsuhiro Honda
Katsunori Shimohara
Katsutoshi Yada

Kazuhiko Tsuda
Kazuhiro Morita
Kazuhisa Seta
Kazumi Nakamatsu
Kenji Matsuura
Koji Harada
Kojiri Tomoko
Kunihiro Yamada
Kyoko Fukuda
Lakhmi Jain
Lars Hildebrand
Lenka Lhotska
Leonardo Franco
Lidia Ogiela
Lily Lin
Lorenzo Magnani
Lorenzo Valerio
Luis Guerrero
M. Howard Williams
Maite Garcia-Sebastian
Maja Vehovec
Mamen Hernandez
Manjunath Joshi
Manuel Grana
Manuel Mora
Marek Kurzynski
Marek Ogiela
Margarita Sordo
Marija Pecina
Mario-Osvin Pavcevic
Mariusz Paradowski
Marko Banek
Marko Rodriguez
Marko Tadic
Masahiro Inuiguchi
Masahiro Tokumitsu
Masao Fuketa
Masayoshi Aritsugi
Masayuki Murakami
Maya Dimitrova
Mayumi Ueda
Mihai Boicu
Miji Jure
Mika Sato-Ilic
Miloslav Vošvrda
Mircea Negoita
Miroslav Karny
Mladen Varga
Mong-Fong Horng
Naohiro Ishii
Naoto Mukai
Natasa Erjavec
Natasa Hoic-Bozic
Ngoc Thanh Nguyen
Nick Younan
Nicola Fanizzi
Nicolai Petkov
Nikica Hlupic
Nina Skorin-Kapov
Nobuhiro Inuzuka
Nobuo Suzuki
Norio Baba
Osamu Fujita
Pablo Roman
Paolo Remagnino
Pascale Kuntz
Paul Brna
Paulo Sampaio
Pedro Morales
Pericles Loucopoulos
Philip Fitch
Rafael Stubs Parpinelli
Renzo Gobbin
Reza Hassanpour
Richard Weber
Robert Gulley
Robert J. Howlett
Ronald Hartung
Rossillawati Sulaiman
Rossitza Setchi
Ruediger Oehlmann
Ryszard Choras
Ryszard Tadeusiewicz
Sanqing Hu
Saric Frane
Satoru Takahashi
Sebastián Ríos
Seiki Akama
Sergio Velastin
Serguei Levachkine
Shih-Wei Sun
Shinji Fukui
Shuming Wang
Shusaku Tsumoto
Shu-Yin Chiang
Sisira Adikari
Sorin Hintea
Stefan Zinsmeister
Stefania Montani
Štěpán Pírko
Takeshi Okamoto
Taki Kanda
Takushi Tanaka
Tamas Sziranyi
Tapio Elomaa
Tatetoshi Ushiama
Tauhid Tayeb
Tetsuhiro Miyahara
Tetsuji Kuboyama
Thomas Breuel
Tohru Matsuodani
Tomislav Smuc
Tomoko Kojiri
Toyohide Watanabe
Tsuyoshi Nakamura
Tuan Pham
Tzung-Pei Hong
Urszula Boryczka
Urszula Markowska-Kaczmar
Ushiama Taketoshi
Valentina Balas
Valentina Zharkova
Vassányi István
Vassilis Kodogiannis
Victor Parada
Waldemar Koczkodaj
Wataru Sunayama
Wen-Li Chyr
Witold Kosinski
Xiaofei Ji
Yen-Wei Chen
Yiannis Kompatsiaris
Yi-Chong Zeng
Yoshinori Adachi

Yuichiro Tateiwa
Yuji Iwahori
Yuji Watanabe
Yuki Hayashi
Yukio Ohsawa
Yumiko Nara
Yu-Ming Liang
Yurie Iribe
Yusuke Hayashi
Zdenko Sonicki
Zeljko Panian
Zhaojie Ju
Zsolt Jankó

KES Conference Series

KES 2009 is part of the KES international conference series.

Conference Series Chairs

L.C. Jain and R.J. Howlett

Sponsors

Millennium Institute of Complex Engineering Systems,
http://www.sistemasdeingenieria.cl/isci/index.php

Millennium Science Initiative, Chilean Government,
http://www.iniciativamilenio.cl/english/index.php

Faculty of Physical Sciences and Mathematics, University of Chile,
http://ingenieria.uchile.cl/

Department of Industrial Engineering, University of Chile,
http://www.dii.uchile.cl/

Table of Contents – Part II

Innovations in Chance Discovery

Advanced Knowledge-Based Systems

Multi-Agent Negotiation and Coordination: Models and Applications

Innovations in Intelligent Systems (I)

Intelligent Technology Approach to Management Engineering

Data Mining and Service Science for Innovation

Knowledge-Based Systems for e-Business

Innovations in Intelligent Systems (II)

Video Surveillance

Social Networks

Advanced Engineering Design Techniques for Adaptive Systems

Knowledge Technology in Learning Support

Advanced Information System for Supporting Personal Activity

Design of Intelligent Society

Knowledge-Based Interface Systems (I)

Knowledge-Based Interface Systems (II)

Knowledge-Based Multi-Criteria Decision Support

Soft Computing Techniques and Their Applications

Immunity-Based Systems

Other Advanced Knowledge-Based Systems (II)

Table of Contents – Part I

Fuzzy and Neuro-Fuzzy Systems

Agent Systems

Knowledge Based and Expert Systems

Other/Misc. Generic Intelligent Systems Topics

Intelligent Vision and Image Processing

Knowledge Management, Ontologies and Data Mining

Web Intelligence, Text and Multimedia Mining and Retrieval

Other Advanced Knowledge-Based Systems (I)

Keynote Speaker Plenary Presentation

Discourse Analysis of Communication Generating Social Creativity

Yoko Nishihara, Yuichi Takahashi, and Yukio Ohsawa

School of Engineering, The University of Tokyo
7-3-1, Hongo, Bunkyo, Tokyo 113-8656, Japan
{nishihara,ohsawa}@sys.t.u-tokyo.ac.jp

Abstract. We have developed a table game named Innovation Game that supports users for thinking up ideas with social creativity. There are two types of players in the Innovation Game, innovators and consumers. While the innovators think up ideas and propose them to the consumers, the consumers criticize the ideas and make decisions whether they buy the ideas or not. In the Innovation Games, the innovators do not only propose their ideas to the consumers, but also improve the ideas using consumer's comments that represent negative impression to the ideas. Therefore, it is considered that ideas with social creativity are related to the negative comments from the consumers. However, the relation between them has not been cleared. In this paper, we analyze discourse texts of communication generating social creativity. The analysis method focuses on the negative comments obtained from the consumers. We analyzed discourse texts of the Innovation Game using the method, and it was verified that the more negative comments the innovators accept, the more ideas with social creativity are obtained.

Keywords: Discourse analysis, Positive/Negative comment, Innovation Game, Social creativity.

1 Introduction

In discussions for making new scenarios, people have to think many hours and exhaust themselves. Since it causes more stress to them, it is difficult for people to exchange their opinions frankly. In order to solve the problem, we have developed a table game named Innovation Game [1].

In the Innovation Game, there are two types of players, innovators and consumers. The innovators think up new business ideas using some cards in which descriptions about existing technologies are written with pictures, and propose the ideas to the consumers. The consumers criticize the ideas and make decisions whether they buy the ideas or not. The innovators do not only propose their ideas, but also improve their ideas referring to comments from the consumers. In communication of the Innovation Game, most of the players are encouraged to give the innovators negative comments that have not been encouraged in the previous brainstorming. Therefore, it is considered that ideas with social creativity are related to the negative comments from the consumers. However, the relation between them has not been cleared.

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 1–8, 2009.

In this paper, we analyze discourse texts of communication generating social creativity. The analysis method focuses on the negative comments obtained from the consumers. The purpose of this paper is to discover how the negative comments from the consumers help the innovators to think up ideas with social creativity.

We define the social creativity as novelty and usefulness for people. In the Innovation Games, the social creativity of idea is evaluated by all of the players considering four measures: idea's cost, idea's utility, idea's reality, and idea's novelty.

2 Related Works

Many researches have been conducted on discourse analysis [2,3,7,8] in computer science. These researches have analyzed discourses with labels called dialogue acts that represent the features of utterance. However, these researches have not focused on negative impressions, and have not prepared such labels. Therefore, it is necessary to prepare a set of labels representing negative impressions in order to analyze the communication of the Innovation Game.

In ethnographic and social-psychological analyses of discourses, the search for features was guided by three analytic concepts characteristic of critical discursive psychology [9]. Though we can obtain the detailed features from discourse texts of the Innovation Game using the previous method manually, we will plan to propose an automatic analysis system for the players of the Innovation Game. Therefore, we do not use the previous method introduced in [9].

Creativity and its support systems have been studied in many fields [4,5,6]. Since the previous methods have supported users for thinking up many ideas not considering the quality of ideas, the ideas have not been made contributions to new businesses and new technologies. We have focused on the social creativity that means novelty and useful for social community. We have developed the Innovation Game as one of the support methods for generating the social creativity.

It is necessary for generating the social creativity to analysis the quality of ideas such as idea's risk, idea's cost, idea's utility, and idea' reality. Since the players are recommended to give the innovators negative comments for the ideas in the Innovation Games. It is useful to discover the relations between the social creativity and the negative comments.

3 Method

All of the procedure is operated manually. A discourse text is transcribed by listening voices of the players. (In the following experiment, we transcribed the discourse texts.) The method takes a discourse text as input. Utterances related to the proposal of ideas are extracted from the discourse text. The discourse text is separated into sub texts that include utterances about one idea. Then labels representing features of utterances are given to each utterance. (In the following experiment, we gave candidates of the label and decided the labels by majority vote of us.) Finally, relations between the ideas with social creativity and the given labels are analyzed.

In the following section, we explain the labels for utterances and the criterion of judging the social creativity of ideas.

3.1 Labels for Utterance

The labels for utterances in a discourse text are the following six ones. We have surveyed the previous researches and the features of utterance in the Innovation Game. Finally, we have constructed a set of labels for analyzing the Innovation Game.

- *Positive*: This label represents positive impression for the last utterance.
- *Neutral*: This label represents non-relation to the last utterance. This also represents an answer for the last question utterance.
- *Negative 1*: This label represents suspicious impression to proposed ideas.
- *Negative 2*: This label represents non-agreement to proposed ideas.
- *Negative 3*: This label represents non-agreement to proposed ideas with questions for innovators.
- *Negative 4*: This label represents non-agreement to proposed ideas with questions for innovators. The consumers point out the weakness of the idea.

3.2 Criterion for Judging the Social Creativity of Ideas

It is considered that the ideas with social creativity are bought by many consumers because such ideas are encouraged in the Innovation Game. Therefore, the criterion for judging the social creativity of ideas is the earnings of each idea from the consumers. We have found correlations between the earnings of ideas and the levels of four measures (idea's cost, idea's utility, idea's reality, and idea's novelty) for evaluating the social creativity of ideas. Therefore, we equaled the ideas with high earnings and the ideas with social creativity.

4 Experiment

We analyzed discourse texts of the Innovation Game using the method. Table 1 shows sets of data used for the experiment. We used two sets of data that are the discourse texts obtained from twice Innovation Games. We instructed the innovators to think up ideas considering social creativity: novelty and usefulness for people. The presented ideas were voted by all of the players considering four measures: idea's cost, idea's utility, idea's reality, and idea's novelty. The consumers were instructed to choose ideas for possessing the ideas with high social creativity. The players are university students, university professors, company employees, and so on. Since they were not always friends, it is considered that a decision bias for idea selection was not much.

Table 2 shows the proposed ideas in the Game 2. Some innovators combined two cards for creating a new idea, and the others combined one card and their own idea proposed in the previous turns (examples are idea O and idea S).

We surveyed the following three relations.

(1) Relation between the social creativity of ideas and the number of utterances.
(2) Relation between the social creativity of ideas and transition of positive/negative labels.
(3) Relation between the social creativity of ideas and the rate of positive labels.

Table 1. Summary of data sets for experiment

	# of innovators	# of consumers	# of utterances	# of proposed ideas
Game 1	3	3	317	16
Game 2	4	3	453	20

Table 2. Proposed ideas and cards used for a new idea in Game 2

Id	Proposed idea	Card 1	Card 2
A	3D-radar for jet plane	3D-rader	Jet Plane
B	Barber service with dancing	Barber	Dancing robot
C	Package design for cup noodle	Photo printer	Mug
D	Control system of human resource	PDO system	RFID system
E	Bath putting bath salt automatically	Unit bath	Pump
F	Robot saving environment	Dancing robot	Eco system
G	Recycle system using a jet plane	Jet plane	Eco system
H	Propeller with carbon heater	Partition	Carbon heater
I	Towel and make up kit	Make kit	Towel
J	Extraordinary sofa	Carbon cloth	Cleaner
K	Electronic ruler	Ruler	Static electricity
L	Wood work kit for children	Microscope	Paper craft
M	Nano-ruler	Ruler	Laser pointer
N	Robot for controlling temperature	Dancing robot	Carbon heater
O	Bath for relaxation	Carbon cloth	J: extraordinary sofa
P	Bath for massaging	Unit bath	Vibrating bed
Q	Health check kit at home	Tongue cleaner	Health check kit
R	Big EL-theater made of glass	Partition	All in one projector
S	Cleaning robot	Dancing robot	K: electronic ruler
T	Bar in tunnel	Map of great spot	Tunnel

5 Results and Discussion

We show the analysis results and discuss the relations between the social creativity of ideas and the negative comments.

5.1 Relation between the Number of Utterances and the Social Creativity of Ideas

Fig. 1 shows the correlation between the number of utterances and the amount of earnings of each idea. The amount of earnings was a product of the price of idea and the number of its sales. The value of correlation was 0.64. The value was high because ideas with the high earnings were obtained after a long communication. In such communications, the innovators and the consumers discussed the weakness of ideas and the innovators improved their ideas using comments from the consumers. Therefore, the correlation between the number of utterances and the social creativity of ideas was obtained. It was verified that there is a relation between the number of utterances and the social creativity of ideas.

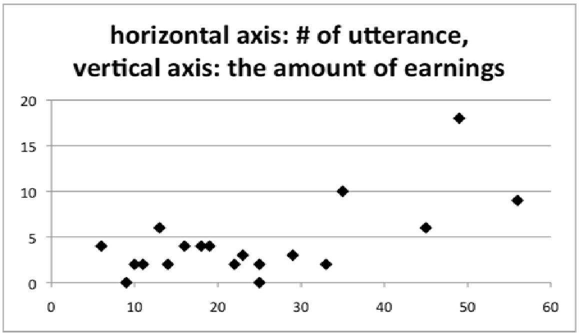


Fig. 1. Relation between the number of utterances and social creativity of ideas

5.2 Relation between the Social Creativity of Ideas and the Transition of Positive/Negative Labels

We surveyed the transitions of labels for utterances. Fig. 2 and Fig. 3 show the transitions of the labels in sub texts. Fig.2 was obtained from a sub text in which the idea obtained high earnings (idea T in Table 2), and Fig. 3 was obtained from a sub text in which the idea obtained low earnings (idea A in Table 2). In Fig. 2 and Fig.3,

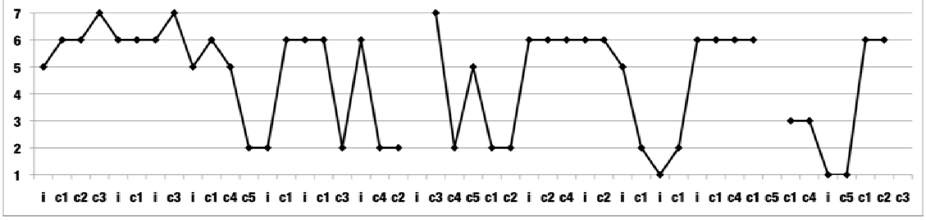


Fig. 2. Transition of the positive/negative labels given to each utterance in a proposal of idea T in Table 2. The horizontal axis denotes the time series. i, c1, c2, and so on denote the players. The vertical axis denotes the label. Number 1, number 2, number 3, and number 4 denote the Negative label. Number 5 denotes the Neutral label. Number 6 denotes the Positive label. Number 7 denotes laugh of players. The idea was about a bar in a tunnel. The idea obtained 10 dollars from the consumers.

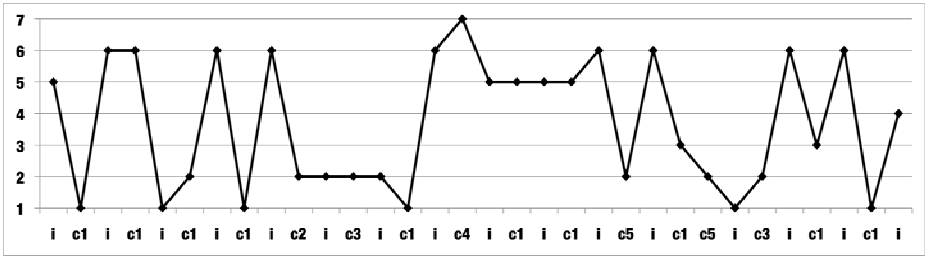


Fig. 3. Transition of the positive/negative labels given to each utterance in a proposal of idea A in Table 2. The idea was about radar of jet plane. The idea obtained 1 dollar from a consumer.

divergences (between the positive label and the negative labels) and convergence (to the positive label and the negative labels) are observed.

In Fig. 2, the transition begins with a divergence from upside to downside, and ends with a convergence to upside. On the other hand, in Fig. 3, the transition begins with a divergence from downside to upside, and does not end with a convergence. From the examples, the transitions are divided with the four features: the divergence, the convergence, upside, and downside. Therefore, we set four patterns as follows:

- Pattern 1: Divergence from upside.
- Pattern 2: Divergence from downside.
- Pattern 3: Convergence to upside.
- Pattern 4: Convergence to downside.

Table 3 shows the results of dividing the transitions. The transitions for ideas obtaining high earnings had two patterns, Pattern 3 after Pattern 1. Table 4 shows an example of the transitions. Pattern 1 is observed from 1st utterance to the 4th utterance and Pattern 3 is observed from 5th utterance to 8th utterance after the innovator (I) agreed with the consumer (C2) at the 4th utterance. It is considered that the Pattern 1 denotes acceptance of the negative comments from consumers, and the Pattern 3 denotes agreement of the social creativity of ideas. The change from Pattern 1 to Pattern 3 was made by the 3rd comment from the consumer C2. The comment improved the idea to be useful and novel for many people. After the 4th utterance from the innovator, the consumers (C2, C3) commented positively to the innovator. Therefore, the idea was bought by many consumers. From the results, it is verified that the convergence to upside is observed after the observation of the divergence between positive labels and negative labels for obtaining ideas with the social creativity.

5.3 Relation between the Social Creativity of Idea and the Rate of Positive Labels

We surveyed the relation between the social creativity of ideas and the rate of positive labels. Table 5 shows the results. The higher earnings the ideas obtained, the more the rate of the positive labels was. It is because the ideas with social creativity are thought up and improved in a comfortable communication between the innovators and the consumers. It is verified that ideas with social creativity are obtained in a frank communication in which positive comments for ideas are uttered.

Table 3. Dividing results of transitions using patterns and the number of ideas. The high earning idea is the idea obtaining less than three dollars. The low earning idea is the idea obtaining more than three dollars.

	# of high earning ideas	# of low earning ideas
Pattern 1: divergence from upside	1	4
Pattern 2: divergence from downside	0	2
Pattern 3: convergence to upside	6	2
Pattern 4: convergence to downside	1	0
Pattern 5: Pattern 3 after Pattern 1	12	2
Pattern 6: including Pattern 5	2	1

Table 4. Example of transitions Pattern 5 (that denotes the transition of Pattern 3 after Pattern 1). C denotes the consumer and I denotes the innovator. The discourse text is obtained a part of idea presentation shown in Fig. 2.

#	Player	Utterance	Label
1	C 1	Why do not use the real tunnel? It is boring only to show the view of tunnel on the screen.	Negative 2
2	I	It is difficult to use the real tunnel for this business, because tunnels are not shown in the center of city.	Negative 2
3	C 2	Would you use the tunnels in dead track? Though you think the tunnel bar only for bar, I think the tunnel bar will become one of the tourist spot like hot spring in Japan.	Negative 2
4	I	I see. That is interesting idea. I agree your comment.	Positive
5	C 2	Do you make one tunnel bar? That is very waste.	Negative 2
6	I	OK. I make tunnel bars as many as possible. I also sell a map for traveling tunnel bars.	Positive
7	C 3	It is very exciting to drink in the real tunnel. If the tunnel bars are built, I certainly visit some tunnel bars in holidays.	Positive
8	I	The safety in tunnels are guaranteed by helmet for construction.	Positive
9		(Players are laughing)	

Table 5. Rate of positive labels for each idea in Game 2

Id	Price	# of sales	# of earnings	Rate of positive comments
A	0	0	0	0.32
B	2	3	6	0.85
C	2	2	4	1.00
D	3	1	3	0.65
E	1	2	2	0.52
F	2	1	2	0.72
G	1	2	2	0.71
H	2	2	4	0.50
I	0	0	0	0.11
J	2	3	6	0.51
K	1	3	3	0.52
L	1	2	2	0.50
M	2	1	2	0.33
N	2	2	4	0.78
O	2	5	10	0.64
P	3	6	18	0.79
Q	1	2	2	0.55
R	2	2	4	0.58
S	2	1	2	0.70
T	3	3	9	0.66

6 Conclusion

In this paper, we analyzed discourse texts of communication generating social creativity. The analysis method focuses on the negative comments.

We have developed a table game named Innovation Game for people to think up ideas with social creativity. In the Innovation Game, negative comments to ideas are encouraged. Therefore, we focused on negative comments to ideas obtained from the consumers in the Innovation Game, and prepared a method for discourse analysis. We used the method and analyzed two data sets of discourse texts that were obtained from the Innovation Games. We discovered that it is important for the innovators to accept the negative comments from the consumers, and to improve their ideas referring to the negative comments.

We will search conditions that encourage the negative comments from the consumers. The conditions are considered that combinations of the players, prices of the proposed ideas, and so on. We will design the environment of communication generating social creativity using the conditions.

References

1. Ohsawa, Y., Maeno, Y., Takaichi, A., Nishihara, Y.: Innovation Game as Workplace for Sensing Values in Product Design and Market. In: Proc. of the 2nd IEEE International Workshop on Data Mining for Design and Marketing 2008, pp. 823–828 (2008)
2. Yamashita, N., Ishida, T.: Analyzing Misconceptions in Multilingual Computer-Mediated Communication. In: Proc. of ACM SIGGROUP Conference on Supporting Group Work, pp. 352–353 (2005)
3. Tokuhisa, R., Terashima, R.: Relationship between Utterances and "Enthusiasm" in Non-task-oriented Conversational Dialogue. In: Proc. of the 7th SIGdial Workshop on Discourse and Dialogue, pp. 161–168 (2006)
4. Amabile, T.M., Conti, R., Coon, H., Lazenby, J., Herron, M.: Assessing the work environment for creativity. Academy of Management Review 39(5), 1154–1184 (1996)
5. Albert, R.S., Runce, M.A.: A History of Research on Creativity. In: Sternberg, R.J. (ed.) Handbook of Creativity. Cambridge University Press, Cambridge (1999)
6. Candy, L., Hori, K.: The digital muse: HCI in support of creativity, Creativity and Cognition, Comes of Age: Towards a New Discipline. ACM Interactions 10(4), 44–54 (2003)
7. Moore, J.D., Pollack, M.E.: A problem for rst: The need for multi-level discourse analysis. Computational Linguistics 18, 537–544 (1992)
8. Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Van, C., Meteer, E.M.: Dialogue act modeling for automatic tagging and recognition of conversational speech. Computational Linguistics 26, 339–373 (2000)
9. Reynolds, J., Wetherell, M.: The Discursive Climate of Singleness: The Consequences for Women's Negotiation of a Single Identity. Feminism & Psychology 13(4), 489–510 (2003)

Value Cognition System as Generalization of Chance Discovery

Yukio Ohsawa

Department of Systems Innovation, School of Engineering
The University of Tokyo
ohsawa@sys.t.u-tokyo.ac.jp

Abstract. Value cognition system (VCS) is a human-centric system to enable value cognition, i.e., sensing, understanding, and taking advantage of latent values of entities. Here, human's talents for value cognition are elevated and activated using tools such as sensors, software for social simulation and data visualization, etc., we will develop newly. The mechanism of this system will be characterized by the spiral process of four phases: (1) *sense*: experience scenes in the real world (2) *recollect*: recollect scenes relevant to a confronted situation, (3) *scenarization*: imagine scenarios to live/work with entities high-lighted via recollection and visualizing the data taken in (1), and (4) *co-elevation*: communicate the imagined scenarios and create a scenario to take advantage of values of entities. The scenario obtained shall be put into action, returning to step (1). Studies on chance discovery so far correspond to VCS applied for transient events.

Keywords: Value, human-machine-environment interaction, physical sensors, human sensors, visualization, meta-cognition, communication, scenarization.

1 Brief Introduction: Value Cognition System

Let us first define *value* as the measure of preference when human makes a decision. Without the cognition of latent values, we cannot create innovative products such as an i-Pod or Electronic cars. For example, the user of i-Pods looks stylish due to the modern outlook and the way of manipulation by the user, and an electronic car enables easy recharging by replacing batteries (on the side of a user) and inexpensive producing (on the side of a car manufacturer). These values, i.e., stylishness, ease, and production cost, came to be introduced in the industrial domains thanks to the human's sense of values rather than solely on the computational technologies for the analysis of business data. The thought of designers in the background of such an innovation is not easy, however: The recent introduction of exchangable batteries is said to have come from the design of a mobile PC, and came to be combined with electronic cars to enable easy and quick recharge of power. This combination came from designers' keen attention to customer opinions, and after all hard discussions about alternative ideas for quickening the power recharge, which enabled their awareness and realization of the new value. We cannot expect all designers can survive the overall process and satisfy the hasty requirement of users, so we need a systematic method

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 9–15, 2009.

for enabling ordinary people to realize value cognition. In this paper, I describe the current vision for creating a Value Cognition System (VCS).

2 The Current State of Sciences and Technologies for VCS

Human and the process for value cognition: Usama Fayyad proposed a clear model of the process of knowledge discovery and data mining in 1995, as summarized in Fig.1 [3]. Here, novel, useful, and non-trivial knowledge is expected to be obtained from data by the spiral of human-machine interaction. Human was positioned as the interpreter of patterns put out from a computer, and also as the evaluator of the novelty, utility, and non-triviality of the acquired knowledge.

On the other hand, Chance Discovery initiated in the early 2000 challenged to discover transitive events significant for the decision making of human. Since 2003 users began to apply methods of chance discovery to real business, and Ohsawa etc noticed it is humans rather than a machine, who run as the driver of the process in successful cases. See Fig.2, which shows the difference of chance discovery from knowledge discovery and data mining in Fig. 1. As shown here, human's spiral of the revised focus on chance events is emphasized [1].

It is noteworthy that Ikujiro Nonaka modeled the process to create not only explicit but also tacit knowledge, in the area of management science, which is called SECI (Socialization, Externalization, Combination, and Internalization) [2]. The lower half of Fig. 2 corresponds to SECI, in that, the process involves the transplantation of tacit knowledge from/to humans via non-verbal communications in workplaces (Socialization), verbal explanation about the tacit knowledge in an explicit representation (Externalization), creating new knowledge by combining pieces of explicit knowledge (Combination), and making of body-fit and context-fit knowledge by putting the new knowledge into action (Internalization). The point of this theory is that the process for creating knowledge can be positioned as a central engine of a company. In other words, the activity to catch latent values in the market, on which to create new strategies for business, can be modeled by the spiral process as SECI, which is another aspect of chance discovery where the understanding of the utility of a chance is desired.

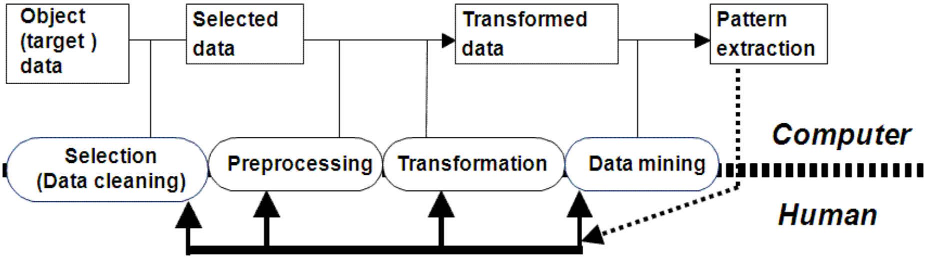


Fig. 1. The process of knowledge discovery and data mining (Fayyad 1995 [3], reformed to compare with Fig.2)

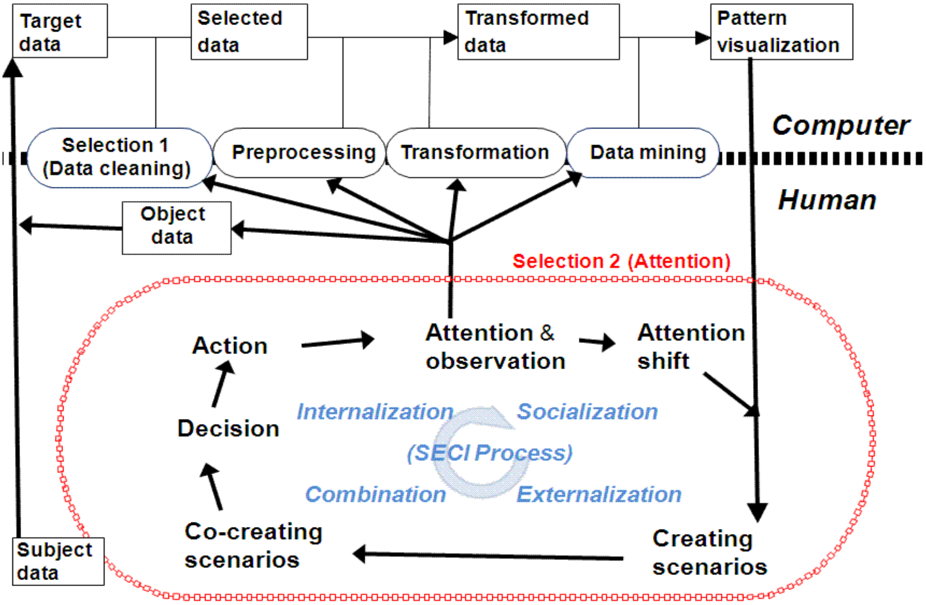


Fig. 2. The relation of processes of KDD, SECI, and Chance Discovery (Ohsawa 2002, reformed)

This point, i.e., the significance of value cognition, still stands in the extension of chance discovery to not only discovering transitive events but also sensing the latent value of entities staying still but not noticed. Donaldson's theory [4] defined value sensing as to feel associated with the content of one's awareness. We can say value cognition is expected to be a spiral process starting from value sensing, to reach the realization of the value via understanding and communication of the value.

3 The Process of Value Cognition

Following the process of chance discovery in Fig.2, we can say now that the fruits of chance discovery went beyond knowledge discovery, in that it enabled the detection of low-probability but useful entities which may be transient and cannot be generalized as knowledge.

However, we have to do at least three significant leaps for realizing a value cognition system as: (1) change human's viewpoint for catching not only transient events and situations, but also objects, people, and cultures which existed long and may embrace a latent value, (2) take care of the huge amount of information in the environment, which have not been included in data (3) develop and integrate tools such as new sensors, tools for training and aiding skills for meta-cognition (cognition of one's own cognition, for noticing the latent criteria of value in the mind) of human(s) , as well as visualization tools as introduced so far in chance discovery, in a way where tools are chosen at suitable timing in the process as in Fig. 3.

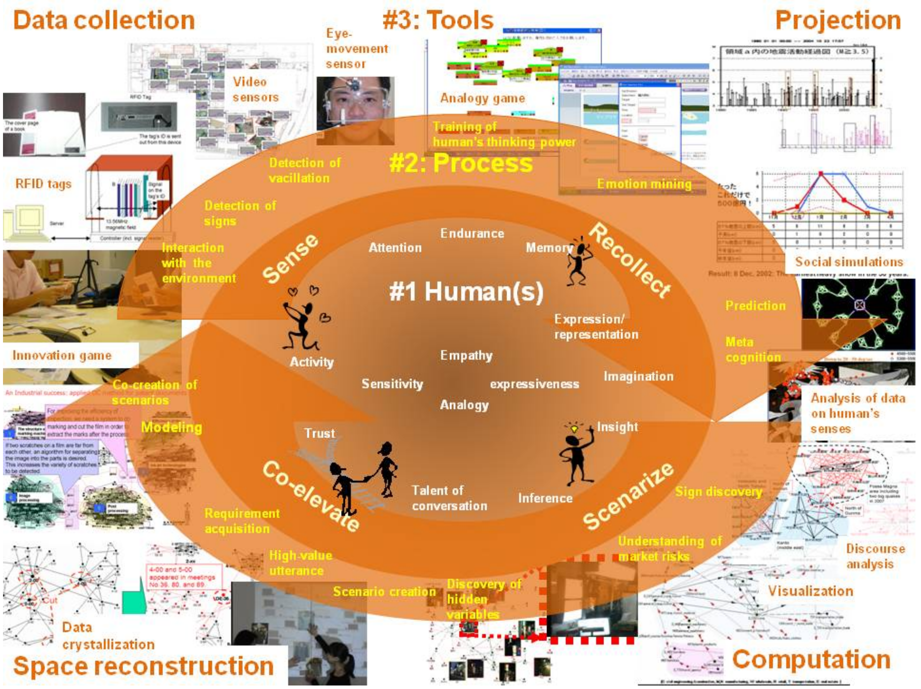


Fig. 3. The hypothetical process of value sensing

These three jumps mean to overcome the problems in the current state of technologies stated in Section 1. That is, for realizing a VCS, we should realize the process of the interaction of decision makers with the environment where the society, nature, and tools (sensing devices, computers, and stationary) surround. This system, having even human as the components, is the VCS composed of three parts:

(1) Human's process of value sensing, where human *sense* entities in the external world in daily life, *recollect* and externalized the experiences, make scenarios (*secenarize*) of the future based on the knowledge acquired from recollected experiences, and communicate with colleagues to *co-elevate* the values of scenarios presented by each other.

(2) Machine's process of data collection via sensing devices, projection to fit the data space so that analysis tools can deal with essential data, computational analysis, visualization, and simulation.

(3) The interaction of (1) and (2), where (a) human's sense and sensing devices collaborate for data collection, (b) human's recollection compensate for the collected data, for the projection of real world to data, (c) human's scenarization is aided by visualization tools, and (d) human's co-elevation is aided by technologies for aiding communication and collaboration, so that the latent values noticed via communication are introduced as variables in the sensing step in following cycles.

4 The Development of Value Cognition System

Based on our experiences in studies on chance discovery for 8 years, we are now aware it is necessary to divide the research activities into subjects of human factors and subjects of developing tools, where the latter are divided in a step-oriented manner, i.e., into teams developing tools to be employed in the steps in the process. Thus, we divide the process into sub-processes, corresponding to research subjects as follows:

Sub-process 1: Elevating human's power*: Here, elevating means both training of human's latent and employment of human's talents in the overall systematic process.*

Subject 1) Environment for elevating human's cognitive power, alents of analogical thinking, insight, empathetic and creative communication will be developed. Methods in the extension of the Analogy Game [5] is expected to contribute to this.

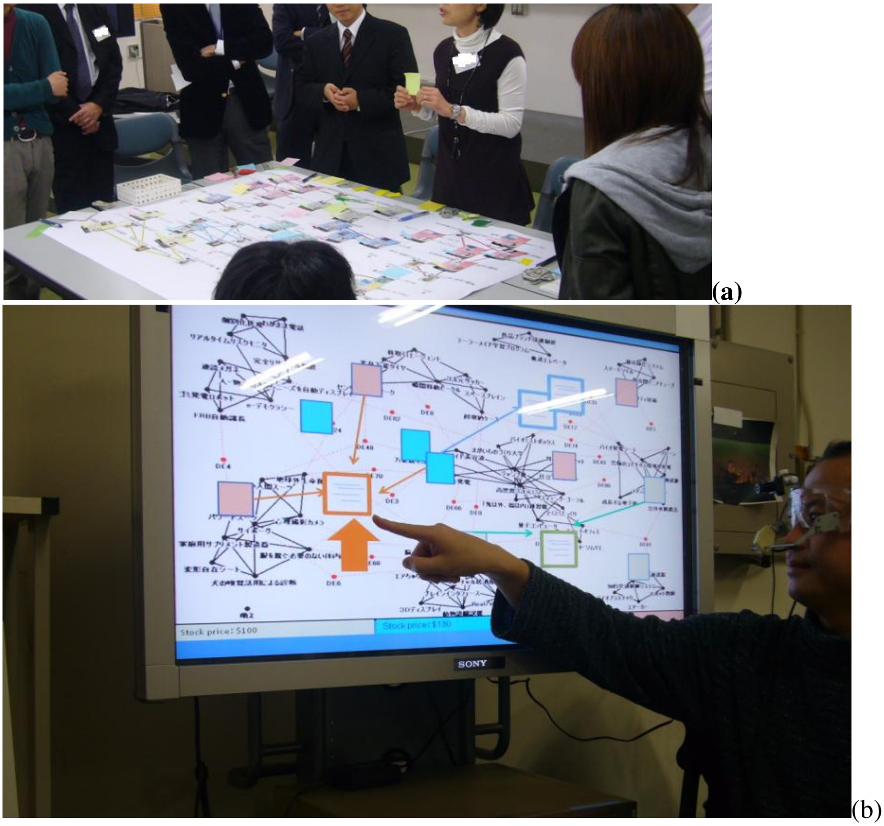


Fig. 4. Innovation game in play (a), and a room in the VCS (b)(just an image)

Subject 2) Methods for elevating human's emotional dimensions, such as sensitivity of the environmental changes, activity, endurance, imagination, empathy, etc. are studied and developed.

***Sub-process 2: Value awareness of individual human aided by computers**: For elevating the steps of recollection and scenarization, methods should be developed for:*

Subject 3) Meta-cognition, where human's consciousness of one's own cognition are mined.

Subject 4) Data-based analysis and visualization of data, and social simulation, so that human can think and discuss about a value gaining/losing scenario in the future.

Subject 5) Discoveries of hidden variables in the real market, in case one notices new value criteria from awareness of the one's own desire.

Sub-process 3: Co-elevation of values

Subject 6) Design of communication environment for elevating the value of integrated scenarios: For example, we can extend the Innovation Game [6] we have been developing so far, to collaborate with methods for extracting and reusing high-value utterances from communication.

Sub-process 4: Humans behavior in and with the real space:

Subject 7) Identification of the target (customers, items, etc) of sensing, in applying the scenario acquired in sub-process 3.

Subject 8) Intelligent sensors: Real time detection of significant signs of value emergence, such as human's and object's noteworthy movements and human's vacillation, using RFID tags, video cameras, etc, based on the model of human behaviors obtained in sub-process 3. Also, the sensor of human cognition such as tracker of human's eye-movements e.g. salesclerk looking at the video of customer's movements, medical doctors looking in a display of patients' image, stock dealers reading news papers, etc, should be developed. However, an essential part of this subject is the software: method to optimization of the sensors' location, and a quick algorithm for deductively inferring human's actions

VCS will work in various workplaces in the real life and businesses of humans. For example, in a supermarket, RFID sensors attached on items have enabled to detect customers' actions. By VCS, the sensitivity of these sensors comes to be controlled by the salesclerk pointing on attention-worthy parts of the store, by a pointing device on the floor map. This decision can be made, based on the meetings of workers in the store introducing the Innovation Game with the game board visualizing the relations between sales items, where new scenarios of sales and customers behaviors are to be obtained. Thus, the sales of products responding to latent customer demands are to be enabled. For example, the salesclerk may re-enforce the sensitivity of RFID sensors at the shelf of cheese, based on the created scenario that in-store customers may increase by showing high-quality cheeses. Then, running the process for value cognition (mentioned in f. Research proposal) again, it comes to be clear that customers desire a combo of cheeses and tea as a gift for aged family. By simulating the market based on the data from test marketing, the sales performance can be predicted and the new project of selling the combo gets started.

5 Conclusions

The cognition of value has been discussed in philosophy and ethics, e.g. Edmond Husserl's value feeling, Max Scheler's value acquisition [7, 8]. Other domain such as design (of products, market, and systems) and service sciences also came to raise discussion relevant to the creation of value. In studies on VCS, we assume no value can be created from true nothing, and focus on the process from the sensing to the realization of the values of events, situations, people, and objects. We regard human's preference in the decisions of daily actions and business as a reflection of value cognition, which makes value (at least indirectly) observable via the decisions of people. This also means that we position the value not only as an attribute of an entity eternal to human, but also as a viewpoint in the mind of human. And, the effects of tools and environments onto human's cognition of values are put into our frame of evidence-based studies on the emergence of value.

As a result, the value sensing system shall increase the total value of the market because it enables to make existing items more valuable for people, as well as to produce new items which are valuable. This goes beyond the traditional approach of management science standing on the basis that values cannot be created without producing new entities with consuming physical resources.

References

1. Ohsawa, Y., McBurney, P. (eds.): Chance Discovery. Springer, Heidelberg (2003)
2. Nonaka, I., Takeuchi, H.: The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation (1995)
3. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: From data mining to knowledge discovery: An overview, Advances of Knowledge Discovery and Data Mining, pp. 1–34. AAAI Press, Menlo Park (1995)
4. Donaldson, M.: Human Minds: An Exploration. Penguin Books, London (1992)
5. Nakamura, J., Ohsawa, Y.: Shift of Mind: Introducing a Concept Creation Model. Information Sciences (in press, 2009)
6. Ohsawa, Y., Maeno, Y., Takaichi, A., Nishihara, Y.: Innovation Game as Workplace for Sensing Values in Design and Market. In: The 3rd International Workshop on Data Mining for Design and Marketing, collocated with IEEE International Conference on Data Mining, Pisa (2008)
7. Scheler, M., Bershady, H.J.: On Feeling, Knowing, and Valuing: Selected Writings (Heritage of Sociology) (1993)
8. Drummond, J.J.: Historical Dictionary of Husserl's Philosophy, Scarecrow Pr. (2008)

Temporal Logic for Modeling Discovery and Logical Uncertainty

Sergey Babenyshev and Vladimir V. Rybakov

Dept. Computing and Math., Manchester Metropolitan University,
Manchester M1 5GD, UK,
V.Rybakov@mmu.ac.uk, S.Babeonyshev@mmu.ac.uk

Abstract. The paper investigates[1] a new temporal logic $\mathcal{LTL}_{\mathcal{DU}}^{\mathcal{Z}}$ combining operations of the linear temporal logic LTL, the operation for discovery and operation for logical uncertainty. Our main aim is to construct a logical framework for modeling logical laws connecting temporal operations and operations of discovery and uncertainty. We consider questions of satisfiability and decidability for $\mathcal{LTL}_{\mathcal{DU}}^{\mathcal{Z}}$. Our principal result is found algorithm which recognizes theorems of $\mathcal{LTL}_{\mathcal{DU}}^{\mathcal{Z}}$ (which implies that $\mathcal{LTL}_{\mathcal{DU}}^{\mathcal{Z}}$ is decidable, and the satisfiability problem for $\mathcal{LTL}_{\mathcal{DU}}^{\mathcal{Z}}$ is solvable).

Keywords: temporal logic, chance discovery, uncertainty, decidability algorithms, Kripke/Hintikka models.

1 Introduction

This paper attempts to embed the notion of Chance for Discovery (CD) in logical framework based on linear temporal logic LTL. CD (cf. Ohsawa and McBurney [10], Abe and Ohsawa [1], Abe et al. [2]) is a contemporary direction in Artificial Intelligence (AI) which analyzes important events with uncertain information, incomplete past data, so to say, *chance* events, where *a chance* is defined as some event which is significant for decision-making in a specified domain. We aim to characterize logical properties of CD within approach based at the linear temporal logic LTL and study interconnections of CD and logical uncertainty.

Temporal logics were first suggested to specify properties of programs in the late 1970's (cf. Pnueli [11]). The most used temporal framework is the linear-time propositional temporal logic LTL, which has been extensively studied from the point of view of various prospects of applications (cf. e.g. Manna and Pnueli [9], Barringer, Fisher, Gabbay and Gough [3]). Model checking for LTL formed a strong direction in logic in computer science, which uses, in particular, applications of automata theory (cf. Vardi [22]). Temporal logics themselves can be considered as a special cases of hybrid logics, e.g. as bimodal logics with some laws imposed on the interaction of modalities to imitate the flow of time.

[1] This research is supported by Engineering and Physical Sciences Research Council (EPSRC), UK, grant EP/F014406/1.

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 16–23, 2009.

The mathematical theory devoted to the study axiomatizations of temporal logics and development of their semantic theory based on Kripke/Hintikka-like models and temporal Boolean algebras formed a highly technical branch in non-classical logics (cf. van Benthem [21,20], Gabbay and Hodkinson [6], Hodkinson [7]). Axiomatizations of various (uni)-temporal linear logics are summarized in de Jongh et al. [5].

At this paper we study a new temporal logic $\mathcal{LTL}^{\mathcal{Z}}_{\mathcal{DU}}$ combining operations of LTL, operation **D** for logical discovery, and operation **Ucn** for logical uncertainty. Motivation for introduction this logic is its high expressive power, which combines the ones from all background logics and, therefore, may model all inherent properties, and, besides, describe possible interaction of combined logical operations. The logic is defined as the set of all formulas valid at all Kripke/Hintikka like models $\mathcal{Z}_C$ based on time indexed by the set $\mathcal{Z}$ of all integer numbers. The problems of satisfiability and decidability for $\mathcal{LTL}^{\mathcal{Z}}_{\mathcal{DU}}$ are of primary interest; we find an algorithm which recognizes theorems of $\mathcal{LTL}^{\mathcal{Z}}_{\mathcal{DU}}$ (which implies that $\mathcal{LTL}^{\mathcal{Z}}_{\mathcal{DU}}$ is decidable, and the satisfiability problem for $\mathcal{LTL}^{\mathcal{Z}}_{\mathcal{DU}}$ is solvable). The algorithm works as follows: an arbitrary formula in the language of $\mathcal{LTL}^{\mathcal{Z}}_{\mathcal{DU}}$ is, first, transformed into a rule in a special normal reduced form, which, then, is checked for validity on special models (of size efficiently bounded by the size of the rule) w.r.t special kind valuations. The general methodology of this paper is borrowed from the paper [18] (and from the research reported recently at the conference KES 2008) where first time logical operation for discovery where first time joined to operations of modal logics Main our aim is to model logical properties of CD and Uncertainty within logical framework of the liner temporal logic LTL. Our results may be classified as theoretical ones concerning mathematical models for CD; they might be useful for researchers interested in logical properties of CD within AI paradigms.

2 Language and Semantics of $\mathcal{LTL}^{\mathcal{Z}}_{\mathcal{DU}}$, Notation

At the beginning we describe sematic structures motivating our introduction of the logical language for $\mathcal{LTL}^{\mathcal{Z}}_{\mathcal{DU}}$. The basic semantic objects upon which we ground our logic are the following Kripe/Hintikka models. A frame

$$\mathcal{Z}_C := \langle \bigcup_{i \in Z} C(i), R, Next, Prev \rangle$$

is a tuple, where Z is the set of all integer numbers, all $C(i)$ are disjoint nonempty sets ($C(i) \cap C(j) = \emptyset$ if $i \neq j$), R is a binary linear relation for time, where

$$\forall a, b \in \bigcup_{i \in Z} C(i)(aRb) \Leftrightarrow [a \in C(i) \& b \in C(j) \& i < j] \vee \exists i \in Z[a, b \in C(i)].$$

The relation $Next$ is defined by $a\ Next\ b \Leftrightarrow [\exists i((a \in C(i)) \& (b \in C(i+1))]$. And, $a\ Prev\ b \Leftrightarrow b\ Next\ a$. The intuitive perception of these frames may be as follows. Any frame $\mathcal{Z}_C$ represents some possible unbounded (in time) computation with multi-possessors as members of $C(i)$; any $i \in Z$ (any integer number i) simulates time tick i, any $C(i)$ consists of processors (computational units) evolved in the computation in time i. Next possible interpretation of models cz_C might be as a representation of a network (web, local net, etc.), where $\mathcal{Z}$ models time ticks (clicks) in future for net navigation; for negative integers from Z, they represent rollback.

To built logical language, we start from the language of LTL, and (to handle possibility for discovery and uncertainty), add the new (unary) logical operations **D** (discoverable) and **Unc** (uncertain). Also, together with the standard language of LTL (with operations **N** (next) and **U** (until)), we will use new binary logical operations $\mathbf{U}_w$ (weak until), $\mathbf{U}_s$ (strong until), **S** (since), $\mathbf{S}_w$ (weak since), $\mathbf{S}_s$ (strong since) and $\mathbf{N^{-1}}$ (previous).

The formation rules for formulas are as usual, and the intended meanings of the operations are as follow:

$\mathbf{D}\varphi$ means the statement φ is *discoverable* in the current state of the current time cluster;

$\mathbf{Unc}\varphi$ means the statement φ is *uncertain* in the current state of the current time cluster;

$\mathbf{N}\varphi$ has the meaning φ holds in the *next time cluster* of states (state);

$\mathbf{N^{-1}}\varphi$ means φ holds in the *previous time cluster* of states (state);

$\varphi\mathbf{U}\psi$ can be read: φ holds until ψ will hold;

$\varphi\mathbf{S}\psi$ φ says that since ψ was true, φ holds until now;

$\varphi\mathbf{U}_w\psi$ has the meaning φ *weakly holds* until ψ will hold;

$\varphi\mathbf{U}_s\psi$ has the meaning φ *strongly holds* until ψ will hold;

$\varphi\mathbf{S}_w\psi$ φ says that since ψ was true, φ weakly holds until now;

$\varphi\mathbf{S}_s\psi$ φ means that since ψ was true, φ strongly holds until now.

For any collection of propositional letters *Prop* and any frame $\mathcal{Z}_C$, a valuation in $\mathcal{Z}_C$ is a mapping which assigns truth values to elements of *Prop* in $\mathcal{Z}_C$. Thus, for any $p \in Prop$, $V(p) \subseteq \mathcal{Z}_C$. We will call $\langle \mathcal{Z}_C, V \rangle$ a model (a Kripke/Hintikka model). For any such model $\mathcal{M}$, the truth values are extended from propositions of *Prop* to arbitrary formulas as follows (for $a \in \mathcal{Z}_C$, we denote $(\mathcal{M}, a) \Vdash_V \varphi$ to say that the formula φ is true at a in $\mathcal{M}_C$ w.r.t. V). The rules are as follows: $\forall p \in Prop,\quad (\mathcal{M}, a) \Vdash_V p \;\Leftrightarrow\; a \in V(p)$; $(\mathcal{M}, a) \Vdash_V \varphi \wedge \psi \Leftrightarrow (\mathcal{M}, a) \Vdash_V \varphi \wedge (\mathcal{M}, a) \Vdash_V \psi$; $(\mathcal{M}, a) \Vdash_V \neg\varphi \Leftrightarrow$ $not[(\mathcal{M}, a) \Vdash_V \varphi]$.

For computation truth values of the logical operation to be discoverable, we apply the rule:

$$(\mathcal{M}, a) \Vdash_V \mathbf{D}\varphi \Leftrightarrow \exists i[a \in C(i) \wedge \exists b \in C(i)(\mathcal{M}, b) \Vdash_V \varphi].$$

Hence, we say φ is *discoverable* at a state of a time cluster $C(i)$ if there is a state in $C(i)$, i.e. in time i, where φ is true. In another words, CD for a statement φ may be satisfied if φ holds at least in one state of the current time cluster.

Further, to compute uncertainty we use:

$$(\mathcal{M}, a) \Vdash_V \mathbf{Unc}\varphi \Leftrightarrow \exists i[a \in C(i) \wedge \exists b \in C(i)(\mathcal{M}, b) \Vdash_V \varphi \wedge$$

$$\exists c \in C(i)(\mathcal{M}, c) \Vdash_V \neg\varphi].$$

That is, we say φ is *uncertain* at a state of a time cluster $C(i)$ if there are two states in $C(i)$, i.e. in time i, where φ is true at one of these states and is false at the another one. This looks as quite plausible way to express uncertainty of φ (though, clearly, one of possible ones, it could be many ways to talk about uncertainty).

Next, we give the rules to compute truth values of logical operations of the linear temporal logic.

$$(\mathcal{M}, a) \Vdash_V \mathbf{N}\varphi \Leftrightarrow \forall b[(a \text{ Next } b) \Rightarrow (\mathcal{M}, b) \Vdash_V \varphi];$$

$$(\mathcal{M}, a) \Vdash_V \mathbf{N^{-1}}\varphi \Leftrightarrow \forall b[(a \text{ Prev } b) \Rightarrow (\mathcal{M}, b) \Vdash_V \varphi];$$

$$(\mathcal{M}, a) \Vdash_V \varphi\mathbf{U}\psi \Leftrightarrow \exists b[(aRb) \wedge ((\mathcal{M}, b) \Vdash_V \psi) \wedge$$

$$\forall c[(aRcRb) \& \neg(bRc) \Rightarrow (\mathcal{M}, c) \Vdash_V \varphi]];$$

$$(\mathcal{M}, a) \Vdash_V \varphi\mathbf{U}_w\psi \Leftrightarrow \exists b[(aRb) \wedge ((\mathcal{M}, b) \Vdash_V \psi) \wedge$$

$$\forall c[(aRcRb) \& \neg(bRc) \& (c \in C(i)) \Rightarrow \exists d \in C(i)(\mathcal{M}, d) \Vdash_V \varphi]];$$

$$(\mathcal{M}, a) \Vdash_V \varphi\mathbf{U}_s\psi \Leftrightarrow \exists b[(aRb) \wedge b \in C(i) \wedge$$

$$\forall c \in C(i)((\mathcal{M}, c) \Vdash_V \psi) \wedge \forall c[(aRcRb) \& \neg(bRc) \Rightarrow (\mathcal{M}, c) \Vdash_V \varphi]];$$

$$(\mathcal{M}, a) \Vdash_V \varphi\mathbf{S}\psi \Leftrightarrow \exists b[(bRa) \wedge ((\mathcal{M}, b) \Vdash_V \psi) \wedge$$

$$\forall c[(bRcRa) \& \neg(cRb) \Rightarrow (\mathcal{M}, c) \Vdash_V \varphi]];$$

$$(\mathcal{M}, a) \Vdash_V \mathbf{S}_w\psi \Leftrightarrow \exists b[(bRa) \wedge ((\mathcal{M}, b) \Vdash_V \psi) \wedge$$

$$\forall c[(bRcRa) \& \neg(cRb) \& (c \in C(i)) \Rightarrow \exists d \in C(i)(\mathcal{M}, d) \Vdash_V \varphi]];$$

$$(\mathcal{M}, a) \Vdash_V \mathbf{S}_s\psi \Leftrightarrow \exists b[(aRb) \wedge b \in C(i) \wedge$$

$$\forall c \in C(i)((\mathcal{M}, c) \Vdash_V \psi) \wedge \forall c[(bRcRa) \& \neg(cRb) \Rightarrow (\mathcal{M}, c) \Vdash_V \varphi]].$$

Given a Kripke structure $\mathcal{M} := \langle \mathcal{Z}_C, V \rangle$ and a formula φ, (i) φ is *satisfiable* in $\mathcal{M}$ (denotation – $\mathcal{M} \Vdash_{Sat} \varphi$) if there is a state b of $\mathcal{M}$ ($b \in \mathcal{Z}_C$) where φ is true: $(\mathcal{M}, b) \Vdash_V \varphi$. (ii) φ is *valid* in $\mathcal{M}$ (denotation – $\mathcal{M} \Vdash \varphi$) if, for any b of $\mathcal{M}$ ($b \in \mathcal{Z}_C$), the formula φ is true at b ($(\mathcal{M}, b) \Vdash_V \varphi$).

For a frame $\mathcal{Z}_C$ and a formula φ, φ is satisfiable in $\mathcal{Z}_C$ (denotation $\mathcal{Z}_C \Vdash_{Sat} \varphi$) if there is a valuation V in the frame $\mathcal{Z}_C$ such that $\langle \mathcal{Z}_C, V \rangle \Vdash_{Sat} \varphi$. φ is valid in $\mathcal{Z}_C$ (notation $\mathcal{Z}_C \Vdash \varphi$) if *not*($\mathcal{Z}_C \Vdash_{Sat} \neg\varphi$).

Definition 1. *The logic $\mathcal{LTL}^{\mathcal{Z}}_{\mathcal{DU}}$ is the set of all formulas which are valid in all frames $\mathcal{Z}_C$.*

We say a formula φ is *satisfiable* iff there is a valuation V in a Kripke frame $\mathcal{Z}_C$ which makes φ satisfiable: $\langle \mathcal{Z}_C, V\rangle \Vdash_{Sat} \varphi$. Clearly, a formula φ is satisfiable iff $\neg\varphi$ is not a theorem of $\mathcal{LTL}^{\mathcal{Z}}_{\mathcal{DU}}$: $\neg\varphi \notin \mathcal{LTL}^{\mathcal{Z}}_{\mathcal{DU}}$, and vice versa, φ is a theorem of $\mathcal{LTL}^{\mathcal{Z}}_{\mathcal{DU}}$ ($\varphi \in \mathcal{LTL}^{\mathcal{Z}}_{\mathcal{DU}}$) if $\neg\varphi$ is not satisfiable. Using the operation $\mathbf{U}$ we, in the well known way, can define all standard modal and temporal operations, e.g. $\Diamond$ (possible), $\Box$ (necessary), $\mathbf{F}$ (*holds eventually*), $\mathbf{G}$ (*holds henceforth*), etc.

The logic $\mathcal{LTL}^{\mathcal{Z}}_{\mathcal{DU}}$, introduced above, is much more expressive compared to standard LTL. We handle variations of the operation Until and Since: the new temporal operations $\mathbf{U}_s$ and $\mathbf{U}_w$ bring new unique features to the language. For instance the formula $\Box_w\varphi := \neg(\top\mathbf{U}_s\neg\varphi)$ codes *weak necessity*, it says that in any future time cluster $C(i)$ there is a state where φ is true. So, this formula way code the *non-vacuity* for a task p computation: in any tick in future at least one possessor unit works with p. The formula $(\neg\varphi\mathbf{U}_w\Box\varphi) \wedge \Diamond\Box\varphi$ codes that, there is a minimal time point i since which φ holds in all states of all future time clusters, but before the time point i the formula φ is false in a state of any time cluster. Such properties are problematic to be expressed in standard modal or temporal operations.

Operations to handle CD and Uncertainty also show general interdependencies in chosen approach. For instance,

$$\mathbf{Unc}\varphi \ \rightarrow\ \mathbf{D}\varphi \ \mathcal{LTL}^{\mathcal{Z}}_{\mathcal{DU}},$$

$$\mathbf{D}\varphi \wedge \mathbf{D}\neg\varphi \ \rightarrow\ \mathbf{Unc}\varphi \ \mathcal{LTL}^{\mathcal{Z}}_{\mathcal{DU}}.$$

3 Key Results, Decidability Algorithm for $\mathcal{LTL}^{\mathcal{Z}}_{\mathcal{DU}}$

For any logical system, one of most fundamental questions is the decidability problem: if there is an algorithm computing theorems of this logic. We address this problem to our logic $\mathcal{LTL}^{\mathcal{Z}}_{\mathcal{DU}}$. The basic technique we use is based on the reduction of formulas in the language of $\mathcal{LTL}^{\mathcal{Z}}_{\mathcal{DU}}$ to special inference rules and the verification of the validity these rules in frames $\mathcal{Z}_C$. This approach uses techniques to handle inference rules from [12] – [19] (where [19] solves decidability of LTL w.r.t. admissibility and again decidability of LTL itself). Recall, a (sequential) (inference) rule is a relation $\mathbf{r} := \frac{\varphi_1(x_1,\ldots,x_n),\ldots,\varphi_l(x_1,\ldots,x_n)}{\psi(x_1,\ldots,x_n)}$, where $\varphi_1(x_1,\ldots,x_n),\ldots,\varphi_l(x_1,\ldots,x_n)$ and $\psi(x_1,\ldots,x_n)$ are formulas constructed out of letters $x_1,\ldots,x_n$. The letters $x_1,\ldots,x_n$ are the variables of $\mathbf{r}$, we use the notation $x_i \in Var(\mathbf{r})$.

Definition 2. *A rule* $\mathbf{r}$ *is said to be* **valid** *in a Kripke model $\langle\mathcal{Z}_C, V\rangle$ (notation $\mathcal{Z}_C \Vdash_V \mathbf{r}$) if $[\forall a\ ((\mathcal{Z}_C, a) \Vdash_V \bigwedge_{1\leq i\leq l}\varphi_i)] \Rightarrow \forall a\ ((\mathcal{Z}_C, a) \Vdash_V \psi)$. Otherwise we say* $\mathbf{r}$ *is* **refuted** *in $\mathcal{Z}_C$, or* **refuted in $\mathcal{Z}_C$ by** *V, and write $\mathcal{Z}_C \nVdash_V \mathbf{r}$. A rule* $\mathbf{r}$ *is* **valid** *in a frame $\mathcal{Z}_C$ (notation $\mathcal{Z}_C \Vdash \mathbf{r}$) if, for any valuation V, $\mathcal{Z}_C \Vdash_V \mathbf{r}$.*

For any formula φ we can convert it into the rule $x \to x/\varphi$ and employ a technique of reduced normal forms for inference rules as follows. Evidently,

Lemma 1. *A formula φ is a theorem of $\mathcal{LTL}^{\mathcal{Z}}_{\mathcal{DU}}$ iff the rule $(x \to x/\varphi)$ is valid in any frame $\mathcal{Z}_C$.*

A rule $\mathbf{r}$ is said to be in *reduced normal form* if $\mathbf{r} = \varepsilon/x_1$ where

$$\varepsilon := \bigvee_{1\leq j\leq l} (\bigwedge_{1\leq i,k\leq n, i\neq k} [x_i^{t(j,i,0)} \wedge (\mathbf{D}x_i)^{t(j,i,1)} \wedge (\mathbf{Unc}x_i)^{t(j,i,2)} \wedge (\mathbf{N}x_i)^{t(j,i,3)} \wedge$$

$$(\mathbf{N^{-1}}x_i)^{t(j,i,4)} \wedge (x_i\mathbf{U}x_k)^{t(j,i,k,0)} \wedge (x_i\mathbf{U}_w x_k)^{t(j,i,k,1)} \wedge (x_i\mathbf{U}_s x_k)^{t(j,i,k,2)} \wedge$$

$$(x_i\mathbf{S}x_k)^{t(j,i,k,3)} \wedge (x_i\mathbf{S}_w x_k)^{t(j,i,k,4)} \wedge (x_i\mathbf{S}_s x_k)^{t(j,i,k,5)}])$$

all x_s are certain letters (variables), $t(j,i,z), t(j,i,k,z) \in \{0,1\}$ and, for any formula α above, $\alpha^0 := \alpha$, $\alpha^1 := \neg\alpha$.

Definition 3. *Given a rule $\mathbf{r_{nf}}$ in reduced normal form, $\mathbf{r_{nf}}$ is said to be a normal reduced form for a rule $\mathbf{r}$ iff, for any frame $\mathcal{Z}_C$, $\mathcal{Z}_C \Vdash \mathbf{r} \Leftrightarrow \mathcal{Z}_C \Vdash \mathbf{r_{nf}}$.*

By following verbatim to Lemma 3.1.3 and Theorem 3.1.11 in [13] we obtain

Theorem 1. *There exists an algorithm running in (single) exponential time, which, for any given rule $\mathbf{r}$, constructs its normal reduced form $\mathbf{r_{nf}}$.*

Decidability of $\mathcal{LTL}^{\mathcal{Z}}_{\mathcal{DU}}$ will follow (by Lemma 1) if we find an algorithm recognizing rules in reduced normal form which are valid in all frames $\mathcal{Z}_C$. We need one more construction on Kripke frames given below. For any frame $\mathcal{Z}_C$ and some integer numbers k_1, m_1, k_2, m_2, where $m_2 > k_2 > k_1 + 3, k_1 > m_1$ we construct the frame $\mathcal{Z}_C(k_1, m_1, k_2, m_2)$ from $\mathcal{Z}_C$ as follows. $\mathcal{Z}_C(k_1, m_1, k_2, m_2) := \langle \bigcup_{m_1 \leq i \leq m_2} C(i), R, Next\rangle$, where R is the accessibility relation from $\mathcal{Z}_C$ extended by pairs (x, y), where $x \in C(i), y \in C(j)$ and $i, j \in [m_1, k_1]$, or $i, j \in [k_2, m_2]$.

Relations $Next$ and $Prev$ are taken from $\mathcal{Z}_C$ and extended by $\forall a \in C(m_2)\forall b \in C(k_2)(a\ Next\ b = true)$; $\forall a \in C(m_2)\forall b \in C(k_2)(b\ Prev\ a = true)$; $\forall a \in C(m_1)\forall b \in C(k_1)(a\ Prev\ b = true)$; $\forall a \in C(m_1)\forall b \in C(k_1)(b\ Next\ a = true)$. For any valuation V of letters from a formula φ in $\mathcal{Z}_C(k_1, m_1, k_2, m_2)$ the truth value of φ can be defined at elements of $\mathcal{Z}_C(k_1, m_1, k_2, m_2)$ by the rules similar to the ones given for the frames $\mathcal{Z}_C$ above (just in accordance with the meaning of logical operations). Due to limitations on the length of the paper we omit a detail description of these rules.

Lemma 2. *A rule $\mathbf{r_{nf}}$ in reduced normal form is refuted in a frame $\mathcal{Z}_C$ iff $\mathbf{r_{nf}}$ can be refuted in a frame $\mathcal{Z}_C(k_1, m_1, k_2, m_2)$ by a valuation V of* special kind, *where the size of the frame $\mathcal{Z}_C(k_1, m_1, k_2, m_2)$ is triple exponential in $\mathbf{r_{nf}}$.*

From Theorem 1, Lemma 1 and Lemma 2 we derive.

Theorem 2. *The logic $\mathcal{LTL}^{\mathcal{Z}}_{\mathcal{DU}}$ is decidable. The algorithm for checking a formula to be a theorem of $\mathcal{LTL}^{\mathcal{Z}}_{\mathcal{DU}}$ consists in verification of validity rules in reduced normal form at frames $\mathcal{Z}_C(k_1, m_1, k_2, m_2)$ of size triple-exponential in the size of reduced normal forms w.r.t. valuations of special kind.*

It is possible also to apply the technique from this paper to weakened versions of the logic $\mathcal{LTL}^{\mathcal{Z}}_{\mathcal{DU}}$, say with omitted strong or weak versions of the operations **U** or **S**, with omitted **N** or $\mathbf{N}^{-1}$ and to obtain similar results about decidability.

4 Conclusion, Future Work

The paper develops a technique to construct mathematical models for investigation of logical operations for CD and Uncertainty. Technically our main aim is to find an algorithm which can compute logical laws of the proposed logic $\mathcal{LTL}^{\mathcal{Z}}_{\mathcal{DU}}$ based on the linear temporal logic LTL. The problem is resolved via model theoretic constructions on special Kripke/Huntikka models and usage of special sequents (inference rules in the reduced form).

There are many prospective avenues of research on logic $\mathcal{LTL}^{\mathcal{Z}}_{\mathcal{DU}}$ and its variants. For instance, the question of finding axiomatization for $\mathcal{LTL}^{\mathcal{Z}}_{\mathcal{DU}}$ is open yet. The problem of computation admissibility of inference rules in $\mathcal{LTL}^{\mathcal{Z}}_{\mathcal{DU}}$ is not investigated. Another interesting problem concerns complexity issues and possible ways of refining the complexity bounds in the algorithm. The problem of describing bases for rules admissible in $\mathcal{LTL}^{\mathcal{Z}}_{\mathcal{DU}}$ logics is also open to date. The results of this paper might be useful for scientists interested in logical properties of CD and Uncertainty and their applications in AI. Tools developed in our paper seem to be promising for investigation properties of logical operations in other similar logics originating in AI.

References

1. Abe, A., Ohsawa, Y. (eds.): Readings in Chance Discovery. International Series on Advanced Intelligence (2005)
2. Abe, A., Hagita, N., Furutani, M., Furutani, Y., Matsuoka, R.: Exceptions as Chance for Computational Chance Discovery. KES Journal 2, 750–757 (2008)
3. Barringer, H., Fisher, M., Gabbay, D., Gough, G.: Advances in Temporal Logic. Applied logic series, vol. 16. Kluwer Academic Publishers, Dordrecht (1999)
4. Crestani, F., Lalmas, M.: Logic and uncertainty in information retrieval. In: Agosti, M., Crestani, F., Pasi, G. (eds.) ESSIR 2000. LNCS, vol. 1980, pp. 179–206. Springer, Heidelberg (2001)
5. de Jongh, D., Veltman, F., Verbrugge, R.: Completeness by construction for tense logics of linear time. In: Troelstra, A.S., Visser, A., van Benthem, J.F.A.K., Veltman, F.J.M.M. (eds.) Liber Amicorum for Dick de Jongh. Institute of Logic, Language and Computation, Amsterdam (2004), `http://www.illc.uva.nl/D65/`
6. Gabbay, D.M., Hodkinson, I.M.: An axiomatisation of the temporal logic with Until and Since over the real numbers. Journal of Logic and Computation 1, 229–260 (1990)

7. Hodkinson, I.: Temporal Logic and Automata. In: Gabbay, D.M., Reynolds, M.A., Finger, M. (eds.) Temporal Logic: Mathematical Foundations and Computational Aspects, vol. 2, pp. 30–72. Clarendon Press, Oxford (2000)
8. Elvang-Goransson, M., Krause, P.J., Fox, J.: Acceptability of arguments as logical uncertainty. In: Moral, S., Kruse, R., Clarke, E. (eds.) ECSQARU 1993. LNCS, vol. 747, pp. 85–90. Springer, Heidelberg (1993)
9. Manna, Z., Pnueli, A.: Temporal Verification of Reactive Systems: Safety. Springer, Heidelberg (1995)
10. Ohsawa, Y., McBurney, P. (eds.): Chance Discovery (Advanced Information Processing). Springer, Heidelberg (2003)
11. Pnueli, A.: The Temporal Logic of Programs. In: Proc. of the 18th Annual Symp. on Foundations of Computer Science, pp. 46–57. IEEE Computer Society Press, Los Alamitos (1977)
12. Rybakov, V.V.: Rules of Inference with Parameters for Intuitionistic logic. Journal of Symbolic Logic 57(3), 912–923 (1992)
13. Rybakov, V.V.: Admissible Logical Inference Rules. Series: Studies in Logic and the Foundations of Mathematics, vol. 136. Elsevier Sci. Publ., North-Holland (1997)
14. Rybakov, V.V.: Construction of an Explicit Basis for Rules Admissible in Modal System S4. Mathematical Logic Quarterly, vol. 47(4), pp. 441–451 (2001)
15. Rybakov, V.V.: Logical Consecutions in Discrete Linear Temporal Logic. Journal of Symbolic Logic 70(4), 1137–1149 (2005)
16. Rybakov, V.V.: Logical Consecutions in Intransitive Temporal Linear Logic of Finite Intervals. Journal of Logic Computation 15(5), 633–657 (2005)
17. Rybakov, V.V.: Until-Since Temporal logic Based on Parallel Time with Common Past. In: Artemov, S., Nerode, A. (eds.) LFCS 2007. LNCS, vol. 4514, pp. 486–497. Springer, Heidelberg (2007)
18. Rybakov, V.V.: Logic of Discovery in Uncertain Situations – Deciding Algorithms. In: Apolloni, B., Howlett, R.J., Jain, L. (eds.) KES 2007, Part II. LNCS (LNAI), vol. 4693, pp. 950–958. Springer, Heidelberg (2007)
19. Rybakov, V.: Linear Temporal Logic with Until and Next, Logical Consecutions, August 2008. Annals of Pure and Applied Logic, vol. 155(1), pp. 32–45 (2008)
20. van Benthem, J., Bergstra, J.A.: Logic of Transition Systems. Journal of Logic, Language and Information 3(4), 247–283 (1994)
21. van Benthem, J.: The Logic of Time. Kluwer, Dordrecht (1991)
22. Vardi, M.: Reasoning about the past with two-way automata. In: Larsen, K.G., Skyum, S., Winskel, G. (eds.) ICALP 1998. LNCS, vol. 1443, pp. 628–641. Springer, Heidelberg (1998)

Evaluation of a Classification Rule Mining Algorithm Based on Secondary Differences

Shusaku Tsumoto and Hidenao Abe

Department of Medical Informatics, Shimane University, School of Medicine
89-1 Enya-cho, Izumo, Shimane 693-8501, Japan
tsumoto@computer.org, abe@med.shimane-u.ac.jp

Abstract. Rule mining is considered as one of the usable mining method in order to obtain valuable knowledge from stored data on database systems. Although many rule mining algorithms have been developed, almost current rule mining algorithms only use primary difference of a criterion to select attribute-value pairs to obtain a rule set to a given dataset. In this paper, we introduce a rule generation method based on secondary differences of two criteria for avoiding the trade-off of coverage and accuracy. Then, we performed an evaluation of the proposed algorithm by using UCI common datasets. In this case study, we compared the predictive accuracies of rule sets learned by our algorithm with that of three representative algorithms. The result shows that our rule mining algorithm can obtain not only accurate rules but also rules with the other features.

Keywords: Rule Mining, Secondary Difference, Rule Evaluation Index.

1 Introduction

In recent years, enormous amounts of data have been stored on information systems in natural science, social science, and business domains. People have been able to obtain valuable knowledge due to the development of information technology. Beside, data mining has been well known for utilizing data stored on database systems. In particular, if-then rules, which are produced by rule mining algorithms, are considered as one of the highly usable and readable outputs of data mining. Considering tradeoff of two criteria when selecting an attribute-value pair for a closure of rules, primary difference is so naive to obtain an adequate volume of rules. Since such rule mining method searches attribute-value space[1] exhaustibly [1], their outputs become enormous number of rules. Considering above mentioned issue, Tsumoto [2] proposed a search strategy to obtain rules, which treat the tradeoff of accuracy and coverage using secondary differences. Therefore, we implemented the idea as a rule mining method. In this paper, we describe the difference between our proposed method and other

[1] The maximum number for n binary attributes is 2^n. Almost of actual datasets have more than 2^n possible rules.

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 24–31, 2009.

representative rule mining method in Section 2. Then, the detail of our method is described in Section 3. In Section 4, we show a result of evaluation by using an implementation of our method. Finally, we conclude this paper in Section 5.

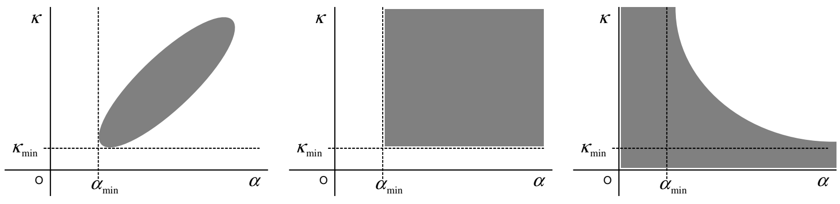
2 Related Work

There are many conventional studies about rule learning algorithms, which are most popular learning algorithms in the machine learning field. As rule mining algorithms, there are the following major approaches: separate-and-conquer [3], methods based on divide-and-conquer, reinforcement learning. Their studies of separate-and-conquer algorithms, which are also called covering algorithms, include many famous algorithms such as AQ family of algorithms [4] and Version Space (VS) [5]. C4.5Rule [6] is based on the decision tree learned with information gain ratio called C4.5, which is classified as the divide-and-conquer approach. Although separate-and-conquer approach has been developed for decades, many new algorithms are developed introducing ideas from the other viewpoints such as APRIORI-C [7]. These algorithms share the following top-level loop: an algorithm searches for a rule that explains a part of its training instances, separates these examples, and recursively conquers the remaining examples by learning more rules until no examples remain. Focusing on the search strategy of rule learning algorithms, they use one simple criterion, such as precision as shown in VS and old AQ family of algorithms. Besides, to treat multiple criteria, other groups of algorithms use combined criterion such as strength of each rule and information gain as shown in Classifier Systems [8], ITRule [9], C4.5 Rule, and PART [10]. There is no algorithm handling two different criteria, because it is a hard work to treat the tradeoff between generality and specificity when an algorithm obtains each rule. APRIORI-C (or so-called predictive Apriori) can use two criteria to search rules from possible rule space. However, they do not treat the tradeoff of generality and correctness of obtained rules, but searching the space exhaustively.

3 A Rule Mining Algorithm Using Secondary Differences of Two Criteria

Tsumoto proposed a rule generation algorithm using secondary differences of two different criteria, α and κ, to generate rules holding both of high accuracy and high coverage. The search space of the algorithm is shown in Fig.1(a) as the gray colored region. Fig.1(b) shows the search space of exhaustive search. To similar, Fig.1(c) shows the search space of the algorithms, which donft consider the tradeoff of generality and correctness of obtained rules.

Fig.2 shows the search strategies of our proposed algorithm. In this figure, $\Delta\alpha(i, i+1)$ and $\Delta\kappa(i, i+1)$ are the primary differences of α and κ. Also, $\Delta^2\alpha(i, i+1, i+2)$ and $\Delta^2\kappa(i, i+1, i+2)$ are the secondary differences of α and κ. For each R, these differences are calculated the following equations to dataset D,



(a) (b) (c)

Fig. 1. Search spaces (gray colored) of the three search strategies

Strategy-I: Selecting rules with $\Delta\alpha > 0 \cap \Delta\kappa = min(\Delta\kappa)$
Strategy-II: Selecting rules with $\Delta^2\alpha = min(\Delta^2\alpha \cap \Delta^2\kappa > 0$

Fig. 2. Search strategies of our rule learning algorithm using secondary differences

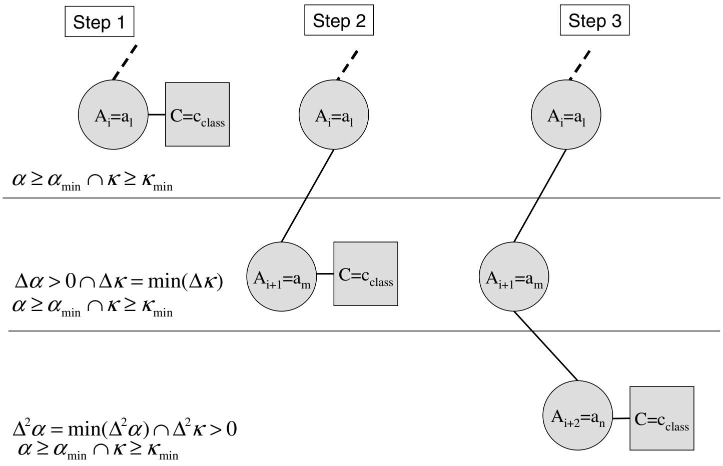

Step 1
Step 2
Step 3
$A_i=a_l$
$C=c_{class}$
$A_i=a_l$
$A_i=a_l$
$\alpha \geq \alpha_{min} \cap \kappa \geq \kappa_{min}$
$\Delta\alpha > 0 \cap \Delta\kappa = \min(\Delta\kappa)$
$\alpha \geq \alpha_{min} \cap \kappa \geq \kappa_{min}$
$A_{i+1}=a_m$
$C=c_{class}$
$A_{i+1}=a_m$
$\Delta^2\alpha = \min(\Delta^2\alpha) \cap \Delta^2\kappa > 0$
$\alpha \geq \alpha_{min} \cap \kappa \geq \kappa_{min}$
$A_{i+2}=a_n$
$C=c_{class}$

Fig. 3. An overview of the steps of the rule generation with the search strategy using secondary differences of two criteria.

where i means the length of the consequents of R. These primary differences and secondary differences are defined as the followings:

$$\Delta\alpha(i, i+1) = \alpha_{R(i+1)}(D) - \alpha_{R(i)}(D) \tag{1}$$

$$\Delta\kappa(i, i+1) = \kappa_{R(i+1)}(D) - \kappa_{R(i)}(D) \tag{2}$$

$$\Delta^2\alpha(i, i+1, i+2) = \Delta\alpha(i+1, i+2) - \Delta\alpha(i, i+1) \tag{3}$$

$$\Delta^2\kappa(i, i+1, i+2) = \Delta\kappa(i+1, i+2) - \Delta\kappa(i, i+1) \tag{4}$$

The overview of the algorithm is shown in Fig.3. Using given two criteria and their lower threshold values, α and κ, by a user, the algorithm firstly obtains the rules with one clause in their consequent. Then, another clause is added to these

```
Input: Dataset,Attributes=$A_{n-1}$+Class, $Alpha_{min}$, $Kappa_{min}$
Output: Ruleset

Begin:
  for(class=0; class<ClassNum; class++){
    for(i=0; i<n-1; i++){
        inclementAntecedent($A_i$, Ruleset, $C_{class}$);
        calculateAlpha(Dataset, Ruleset);
        calculateKappa(Dataset, Ruleset);
        selectRules(Ruleset, =>Alphamin);
        selectRules(Ruleset, =>Kappamin);
        for(j=i+1; j<n-1; j++){
          inclementAntecedent($A_j$, Ruleset, $C_{class}$);
          calculateAlpha(Dataset, Ruleset);
          calculateKappa(Dataset, Ruleset);
          selectDelta(Ruleset, Alpha,>0);
          selectDelta(Ruleset, Kappa, min);
          selectRules(Ruleset, =>Alphamin);
          selectRules(Ruleset, =>Kappamin);
          for(k=i+2; k<n-1; k++){
             inclementAntecedent($A_k$, Ruleset, $C_{class}$);
             calculateAlpha(Dataset, Ruleset);
             calculateKappa(Dataset, Ruleset);
             selectDelta2(Ruleset, Alpha, min);
             selectDelta2(Ruleset, Kappa, >0);
             selectRules(Ruleset, =>Alphamin);
             selectRules(Ruleset, =>Kappamin);

          }
        }
     }
  }
End;
```

Fig. 4. Pseudo code of the rule learning algorithm using secondary differences

rules. The rules, which donft satisfy the strategy-I in Figure 2, are pruned. The remaining rules are stored in to the rule set, and go to next step. In the next step, the rules are added another clause again. Then, the rules, which donft satisfy the strategy-II, are pruned. Storing rules, which satisfy α_{min} and κ_{min}, on each step, the algorithm iterates these steps for each attribute $i(i = 1, 2, ..., A - 2)$[2] and class value. Fig.4 shows a pseudo code of this algorithm.

4 Evaluation by Using UCI Common Datasets

In this section, we describe about a case study of an implementation of the algorithm explained in Section 3. We implemented the algorithm in Java, combining a rule evaluation index calculation module called COIN [11]. Using the implementation, we generated rule sets to the five datasets from UCI Machine Learning Repository [12]. Table1 shows the descriptions of the nine datasets that are used in this evaluation.

The numerical attributes in these datasets, we discretized each attributes into ten bins with equalized width. For example, the number of possible rules

[2] A is the number of attributes in a dataset D.

Table 1. Description of the nine UCI datasets

Datasets	# of Att.	# of Class	Size
iris	4	3	150
labor	16	2	57
breast-cancer	9	2	286
hepatitis	19	2	155
heart-statlog	13	2	270
balance-scale	4	3	625
lymph	18	4	148
glass	9	6	214
diabetes	8	2	768

Table 2. The lower threshold values of the nine UCI datasets obtained by 1000 time bootstrap iterations for PART rule sets

Dataset	Precision	Recall
iris	0.43	0.09
labor	0.54	0.27
breast-cancer	0.00	0.00
hepatitis	0.01	0.01
heart-statlog	0.00	0.00
balance-scale	0.00	0.00
lymph	0.00	0.00
glass	0.00	0.00
diabetes	0.00	0.00

of iris, which has four numerical attributes, is $10^4 \times 3$. Then, the accuracies of the rule sets are compared with that of OneR [13], PART, and unpruned J4.8, which are implemented in Weka [14]. In this experiment, we specified precision and recall to search for rule sets. Precision shows the correct rate of the prediction of each rule as shown in Equation 5. In similar, recall shows the rate of correctly predicted instances in the dataset D for each class, as shown in Equation 6.

$$Precision_R = P(D|R) \tag{5}$$

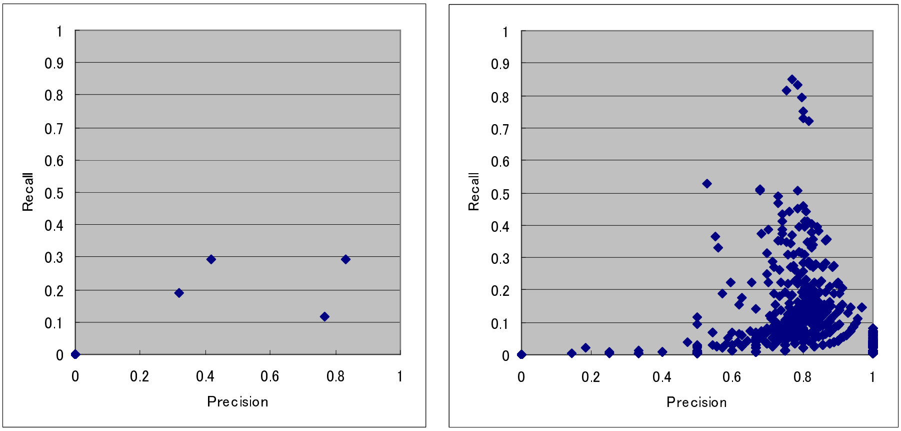
$$Recall_R = P(R|D) \tag{6}$$

Our rule learning method also needs lower thresholds, α_{min} and κ_{min}. We set up these lower thresholds as $Precision_{min} = 0.5$ and $Recall_{min} = 0.3$ as 'Setting1'. As for more comparable lower thresholds, we used the averaged lower values of the rule sets obtained by PART. These values are gathered by 1000 time's bootstrap iterations of PART for each dataset. The threshold values of 'Setting 2' are shown in Table2.

Table3 shows the averaged accuracies of each algorithm to the nine datasets. These accuracies are obtained with 100 times repeated 10-fold cross validation. The highest accuracies for each dataset are emphasized as the oblique numbers.

Table 3. The average accuracies (%) of the four rule learning algorithms and standard deviations (SDs) of the accuracies

Dataset	Proposed Method (Setting 1)		Proposed Method (Setting 2)		OneR		PART		J4.8(unpruned)	
	Acc.	SD	Acc.	SD	Acc.	SD	Acc.	SD	Acc.	SD
iris	77.3	12.1	76.0	13.4	96.0	4.7	94.8	5.5	*96.0*	4.7
labor	75.4	20.2	*76.7*	20.5	68.9	14.8	71.2	12.5	68.6	11.7
breast-cancer	71.3	9.1	55.3	15.8	67.1	5.9	69.7	7.1	*73.9*	5.6
hepatitis	79.5	11.3	72.6	12.0	*83.1*	8.2	80.3	8.6	82.8	8.0
heart-statlog	*80.7*	6.2	64.7	8.5	71.6	7.8	79.1	7.6	80.3	7.7
balance-scale	69.5	5.6	44.1	18.0	57.7	3.5	*76.3*	4.7	64.5	4.6
lymph	78.1	9.5	59.5	7.6	74.6	10.6	*80.6*	9.7	78.6	10.0
glass	44.1	10.8	41.8	14.2	50.8	9.5	55.7	9.5	*57.6*	8.8
diabetes	65.1	3.1	65.1	4.6	*74.4*	4.3	73.3	4.8	74.0	4.0



(a) (b)

Fig. 5. Scatter plots of the obtained rules to Precision-Recall plane. (a):the rules obtained by PART (n=19). (b):the rules obtained by our rule mining algorithm (n=1417)

Although J4.8 achieved most accurate result for these datasets, our rule mining algorithm achieved the highest accuracies for 'labor' and 'heart-statlog' datasets. This result indicates that our rule mining method can outperform to other accuracy oriented rule mining algorithms, when an adequate minimum thresholds are given.

The disadvantage of the accuracies is caused by the two major reasons: lower threshold values and the confliction avoid strategy. The given lower threshold values were not optimized to obtain accurate rule sets. In addition, we avoid conflictions of rules using "better precision first", when predicting the class for each test instance. The strategy should be selected an adequate one to predict test instances more correctly.

Focusing on the feature of the obtained rules, Fig. 5 shows the scatter plots of obtained rules for 'breast-cancer' dataset by using PART and our proposed algorithm with Setting 1. Although the minimum thresholds for our method are

$Precision_{min} = 0.00$ and $Recall_{min} = 0.00$, the rules have more high *Precision* and *Recall* values. The shape of the plot of Fig.5 (b) shows trajectories of the search iterations from initial rules that have only one condition in their antecedent to the rules with longer antecedent.

5 Conclusion and Future Work

In this paper, we described a rule mining algorithm using secondary difference of two objective rule evaluation indices. The result of the case study in Section 4 shows that our proposed algorithm can obtain the rule sets with different features, comparing with the four representative rule learning algorithms. The differences appeared not as the correctness of rule sets, but as rules with both high generality and high accuracy. In the future, we will evaluate usefulness of the proposed method with actual medical data, comparing with the other metrics. Then, we will also obtain rule sets with pairs of objective rule evaluation indices, which have different functional behaviors [15].

References

1. Furnkranz, J., Flach, P.A.: ROC 'n' rule learning: towards a better understanding of covering algorithms. Machine Learning 58(1), 39–77 (2005)
2. Tsumoto, S.: Accuracy and coverage in rough set rule induction. In: 11th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (2006)
3. Furnkranz, J.: Separate-and-Conquer Rule Learning. Artificial Intelligence Review 13(1), 3–54 (1999)
4. Michalski, R.S.: On the QuasiMinimal Solution of the Covering Problem. In: Proceedings of the 5th International Symposium on Information Processing (FCIP69), vol. A3(Switching Circuits), pp. 125–128 (1969)
5. Mitchell, T.M.: Generalization as Search. Artificial Intelligence 18(2), 203–226 (1982)
6. Quinlan, J.R.: Programs for Machine Learning. Morgan Kaufmann, San Francisco (1992)
7. Jovanoski, V., Lavrac, N.: Classification Rule Learning with APRIORI-C. In: Brazdil, P.B., Jorge, A.M. (eds.) EPIA 2001. LNCS (LNAI), vol. 2258, pp. 44–51. Springer, Heidelberg (2001)
8. Booker, L.B., Holland, J.H., Goldberg, D.E.: Classifier Systems and Genetic Algorithms. Artificial Intelligence 40, 235–282 (1989)
9. Goodman, R.M., Smyth, P.: The induction of probabilistic rule sets—the Itrule algorithm. In: Proceedings of the sixth international workshop on Machine Learning, pp. 129–132 (1989)
10. Frank, E., Witten, I.H.: Generating accurate rule sets without global optimization. In: The Fifteenth International Conference on Machine Learning, pp. 144–151 (1998)
11. COIN Project, `http://coin.sourceforge.jp/`

12. Asuncion, A., Newman, D.J.: UCI Machine Learning Repository. University of California, School of Information and Computer Science, Irvine (2007), http://www.ics.uci.edu/~mlearn/MLRepository.html
13. Holte, R.C.: Very simple classification rules perform well on most commonly used datasets. Machine Learning 11, 63–91 (1993)
14. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, San Francisco (2000)
15. Abe, H., Tsumoto, S.: Analyzing behavior of objective rule evaluation indices based on a correlation coefficient. In: Lovrek, I., Howlett, R.J., Jain, L.C. (eds.) KES 2008, Part II. LNCS (LNAI), vol. 5178, pp. 758–765. Springer, Heidelberg (2008)

Communication between Living and Scientific Knowledge as Chance Discovery

Yumiko Nara

The Open University of Japan
2–11 Wakaba, Mihama-ku Chiba City 261–8586 Japan
narayumi@u-air.ac.jp

Abstracts. Knowledge of living means information that people have obtained from their daily lives, and skills and wisdom that they have acquired through everyday experience or tradition. This paper aims to clarify three issues: (1) the uniqueness of living and scientific knowledge, (2) the significance of communication in these two kinds of knowledge, (3) the potentiality of a double helix structure for the two types knowledge in chance discovery, i.e. collaboration between lay subjects and specialists. These tasks were approached through theoretical and empirical research using concrete data obtained from a questionnaire and a case study. As a result, specialists and lay subjects were found to have outstanding knowledge in mutually different contexts even though they had limitations; solutions to problems were obtained through collaboration using mutual knowledge that was obtained on an equal footing.

Keywords: knowledge of living, scientific knowledge, lay subjects, specialists, communication, collaboration.

1 Introduction

We spend our everyday lives with various and abundant knowledge gained from living. Knowledge of living (LK) consists of information that people have obtained from their daily lives, and skills and wisdom that they have acquired through everyday experience or tradition. Scientific knowledge (SK), in this paper, is defined as knowledge gained though working in science.

LK and SK seem to have been treated separately in the fields of science and policy making. They might have been adapted for modern and present-day societies to deal with issues separately, and developed by promoting separation from individual standpoints such as the views of teachers and students, specialists and citizens, and producers and consumers.

We have recently been exposed to more complicated problems in society. Moreover, the stakeholders have also been diversified. It is therefore necessary to find new ideas and methods of solving problems related to various types of stakeholders and communication between various values, various standpoints, and different kinds of knowledge is now needed. This communication seems to have become especially important in solving problems such as those concerning the environment, consumers, communities, security, and safety.

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 32–41, 2009.

There are three main aspects in the relationships between LK and SK: (1) Lay subjects include existing SK in their daily lives. (2) Lay subjects working in science or specialists positively implement SK that is needed in their daily lives, even in helping to create new SK. (3) Lay subjects contribute to solving problems using their LK equally with SK.

Ordinary lay subjects in these three aspects are regarded as "trustworthy participants" within the framework of solving problems who use their LK as one form of chance discovery. This can be expected to function as double helical processing between LK and SK. The double helix model, for the process whereby we discover a chance (as a new opportunity), was conceptualized by Ohsawa (Ohsawa 2001; 2003) and has been applied and developed in various kinds of cases. Subjective data (human) and objective data (from data mining systems such as KeyGraph) work together, each progressing spirally toward finding and creatively reconstructing ideas. It is difficult to obtain chance neither by only machinery view nor by only human view; same relationship could be applied to SK and LK.

Based on the above, this paper discusses three issues: (1) the uniqueness of LK and SK, (2) the significance of communication between LK and SK, and (3) the form and potentiality of the double helix model in chance discovery, i.e., the cooperation of LK (ordinary lay subjects) and SK (specialists) with concrete data obtained from a case study and a questionnaire.

2 Approach toward Knowledge of Living and Scientific Knowledge

2.1 Studies on and Practice in Relationship between Science and Technology (S&T) and Society

Japan has been entering a new era as a knowledge-based society. The subsystems in this society, such as industry, academia, government, and citizens are under great pressure to change with this innovation. According to The Third Science and Technology Basic Plan of the Japanese Government (March 2006), innovation is described as advanced scientific findings and technical inventions combined with human insights that evolve and generate new ideas, methods, or devices that are valuable to society and the economy.

The role and impact of science and technology (S&T) have also been evolving in a complex manner. Our lives have certainly been enriched by access to abundant knowledge and technologies. Some of these, however, despite being developed and used with the best of intentions, have unintended consequences, creating major social problems. As a result, scientific and technological developments are a growing source of concern even though they are essential parts of our society. Under such circumstances, we must not only take advantage of S&T but also construct a new relationship between them and society based on various values at the individual, local, national, and global levels.

Science and technology is the main research area where such problems are explored. The Japanese Society for Science and Technology Studies (JSSTS) was established with the aim of providing a forum for genuine cross-disciplinary, critical, and constructive academic studies into questions related to S&T and society. There are also international societies such as the Society for Social Studies of Science (established in 1976) and the European Association for the Studies of Science and Technology (in 1981). Science and technology for society involves solving social problems, and managing society smoothly (Horii 2006). The Research Institute of Science and Technology for Society (RISTEX), for instance, promotes practical research, and formulates networks with various stakeholders to solve social problems.

There have been significant approaches to communication between different types of knowledge. Communications between specialists and citizens are called "science communication" or "science and technology communications"; some studies that have tried to implement these have also been conducted in Japan. Support by the government given to science and technology communications began in 2005. There are also some programs such as the "Science and Technology Interpreter Training Program (The University of Tokyo)" and the "Communications Design Center (Osaka University). "Science cafés" were held at 21 locations initiated by the Japan Society for the Promotion of Science in 2006, and their number has been increasing throughout the country. Research on skills to promote S&T communications at educational sites (primary through to higher education) has also been done (Chiba 2007).

2.2 Conceptualization and Uniqueness of LK and SK

Science involves work to acquire systematic knowledge based on certain rules, which is called scientific methodology. A certain kind of objectivity and rationality are guaranteed by SK, which enables close examination with scientific methodology. However, LK has a certain subjectivity and irrationality in the sense that it has not been obtained though examination with scientific methodology. This is one of its most typical peculiarities, which is in contrast to SK.

Science tries to approach problems from the viewpoint of researching and understanding objects; fields related to living such as policy of science are not exceptions either. The problems are arranged from a viewpoint, which the object should be examined from. The problems for LK, on the other hand, appear at the actual living site. The knowledge that is used to understand and solve the problems is also created at the actual living site. The problems are understood from the standpoints of individuals who actually live there.

SK is unique in three respects: (1) As SK has demonstrated its strong influence in contributing to modern civilization and society, it tends to be treated as entire knowledge. (2) Only things that are systematized and shared by using scientific methodology are included in SK. (3) The ranges of problems being investigated by scientific methodology is expanding, and contains numerous things. SK is subdivided, specialized, and it is very difficult to look at the whole.

LK has three features: (1) LK is important knowledge and wisdom concerning one's life experienced by all people. They cannot live without it. (2) LK is based on individual experience. It is difficult to generalize one's LK to others. It is also difficult to express one's LK with words. Consequently, LK tends to be personal, and is seldom shared among others. (3) As daily life consists of overall activities (working, consuming, nursing, child raising, cooking, resting, etc.) and elements (monetary resource, time, space, interpersonal resource, etc.) that maintain one's life as it adjusts to environments, LK is also formulated with overall elements related to such activities. Moreover, one individual understands the whole.

2.3 Significance of Communication between LK and SK

It is difficult to cover both forms of knowledge within the same context when these differences in features are singled out. LK and SK have actually been treated separately, especially in modern society. The whole of society is currently structured by extremely detailed sub-systems that divide work and activities. It has become more difficult to understand or assess others' work or activities. It is not easy for non-specialists to access and understand the work of specialists. Even when specialists and non-specialists discuss the same problem, LK is not easily understood by the former. The relation between specialists and lay people is that between people who teach and those who are taught. The more SK has been produced, the greater the separation between specialists and lay people has become.

There is a model of communication that characterizes this phenomenon called the “deficit model”, which is based on the idea that as citizens lack knowledge about science and technology, it is important for specialists to give them accurate knowledge that is easy for them to understand. According to the deficit model, the reason ordinary people feel anxious and are opposed to social problems such as the acceptance of nuclear power plants is only because they are ignorant about science and technology. If accurate knowledge were given, unease and repulsion would be reversed, and people would accept them without any emotional fuss. The “Public Understanding of Science: PUS” was advocated based on this idea. However, the deficit model actually came to be criticized strongly in the mid 1990s; it was too simple to enable the views and actions of people to be understood who were encountering social problems at real sites such as those experiencing bovine spongiform encephalopathy (BSE) or genetically modified products (Wynne 1995. As a result, interactive conversation and communications came to be more valued and a movement involving "Public Engagement with Science and Technology” (PEST) appeared (Hirakawa 2003).

Furthermore, specialists and lay people have outstanding knowledge in mutually different contexts even though they have mutual limitations. The possibility of approaching and solving problems in our complex world might occur if they faced each other on an equal footing, and used their mutual knowledge to complement that of each other. The communications between the two kinds of knowledge, i.e., LK and SK, may fulfill such expectations.

3 Examination into Potentiality of Communication between LK and SK

3.1 Questionnaire to Lay Subjects

This section discusses an examination into the potentiality of communication between LK and SK assessed with some empirical data. First, the views of lay and specialist subjects toward LK and SK as well as their relationships had to be obtained. To clarify this, the author conducted a social survey with a questionnaire.

Index of Main Variables

The 10 main statements (a–j) subjects had to respond to were:

a. Scientific developments have more of a positive effect on human beings than negative.
b. Scientific developments have contributed enormously to my daily life.
c. Statements by scientists and professionals are credible.
d. There are still many things in life and society that cannot be solved, even by applying scientific knowledge.
e. Knowledge acquired through actual experience is in fact more important to human beings than that acquired through scientific methods.
f. Knowledge that is useful in daily life and that used in science are quite different.
g. Scientists and professionals should value the knowledge from the general public or ordinary citizens and apply that to their studies.
h. The general public or ordinary citizens cannot argue with what scientists and professionals say.
i. Statements by scientists and professionals are too difficult to comprehend.
j. Statements by scientists and professionals have had no effect on my life.

They were asked the same question for each of the statements, i.e., "How much do you agree with the following statements? ('Strongly Agree', 'Somewhat Agree', 'Somewhat Disagree', or 'Strongly Disagree') Circle one for each statement."

Outline of Survey

This survey was carried out within the following six-part framework: (1) The population and subjects were males and females who were 20–69 years old and lived throughout Japan. (2) The questionnaire was returned by mail, and (3) the sampling ledger was the NOS list. In the entire country, subjects were sampled randomly according to sex, age, and population percentage. (4) The number of useable samples was 1,050. (5) The survey period was February 13–29, 2008. (6) The organization implementing the investigation was the Nippon Research Center. The basic attributes of the respondents were as follows. Females comprised 54.6% of the subject and males 45.4%. Ages (average 45.65 years old) ranged from 20–29, (14.1%), from 30–39 (23.5%), from 40–49 (21.1%), from 50–59 (20.3%), and from 60–69 (20.9%).

Status Quo of Views by Lay Subjects toward LK and SK

Only the results for frequencies of variables have been listed in Table 1 due to space limitations.

The results revealed that lay subjects accepted SK as essential to improve their quality of life. At the same time, almost 80% of respondents felt that the statements by scientists and professionals were too difficult to comprehend; this reveals how necessary and important it is for specialists to give lay people accurate knowledge in simple form that they can clearly understand. If just this finding were focused on, it would remain as PUS based on the idea of the deficit model.

However, we had to focus on other reactions by respondents, i.e., the confidence they had in their LK. They thought that knowledge acquired through actual experience was in fact more important to human beings than that acquired through scientific methods, and they even felt that scientists and professionals should value the knowledge obtained from the general public or ordinary citizens and apply this to their

Table 1. Views of lay subjects toward LK and SK

(%)

	Strongly Agree	Somewhat Agree	Somewhat Disagree	Strongly Disagree
a. Scientific developments have more of a positive effect on human beings than a negative one.	12.0	58.8	26.3	2.8
b. Scientific developments have contributed enormously to my daily life.	16.1	64.1	18.4	1.4
c. Statements by scientists and professionals are credible.	3.4	56.0	37.5	3.0
d. There are still many things in life and society that cannot be solved, even by applying scientific knowledge.	38.4	53.2	7.8	0.6
e. Knowledge acquired through actual experience is in fact more important to human beings than that acquired through scientific methods.	16.4	58.1	24.9	0.7
f. Knowledge that is useful in daily life and that used in science are quite different things	17.0	55.0	26.0	2.0
g. Scientists and professionals should value the knowledge from the general public or ordinary citizens and apply that to their studies.	17.8	60.8	19.0	2.4
h. The general public or ordinary citizens cannot argue with what scientists and professionals say.	4.8	18.4	58.6	18.3
i. Statements by scientists and professionals are too difficult to comprehend.	16.6	53.2	27.3	2.9
j. Statements by scientists and professionals had no effect on my life.	3.6	25.7	59.5	11.3

studies. This finding indicates the possibility of germinating PEST and communications between LK and SK.

3.2 Case Study on Crime-Prevention Activities in Sakai-City

This section discusses an examination into the potentiality of communication between LK and SK using a case study.

The environment around children has recently been changing and serious crimes have been committed against them as victims. Approaches that have defended children against crime have been taken by families, schools, local communities, and NPOs. However, many of these have been conducted individually and experientially, and have not been based on tangible data or scientific methods. Then, it had to be recognized that people who work on this problem introduce scientific findings and methods into the crime-busting measure while receiving the specialists' co-working, in order to solve social problems effectively and continuously. A case is discussed here that was aimed at achieving this goal, i.e., crime-prevention activities in the city of Sakai to protect children.

Framework of Activities

Figure 1 shows the stakeholders who joined in these activities. The main lay subjects were residents of the city of Sakai (Tomioka district) and members of the NPO Sakai hill-front forum. This NPO is an organization that consists of residents of the Tomioka district. The specialists were experts in crime prevention such as police, and engineers who developed an Information and Communication Technology (ICT) system,

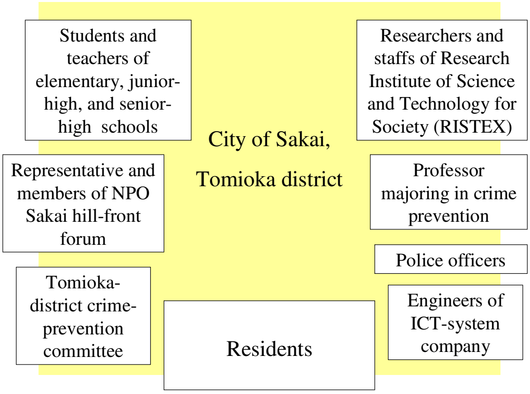

Students and teachers of elementary, junior-high, and senior-high schools
Researchers and staffs of Research Institute of Science and Technology for Society (RISTEX)
City of Sakai,
Tomioka district
Representative and members of NPO Sakai hill-front forum
Professor majoring in crime prevention
Police officers
Tomioka-district crime-prevention committee
Residents
Engineers of ICT-system company

Fig. 1. Stakeholders in crime-prevention activities in Sakai

and university professors who cooperated in these activities. Within this framework, RISTEX (Chap.2.1) also played an important role as it aimed to induce and support the types of innovation that addressed the needs of the public, such as ensuring the safety and security of residents and increasing their quality of life.

The purpose of these activities was to "alleviate residents' concerns and create a safe regional society in which crimes against children would not be tolerated to construct a new social system." The three main pillars of the activities to achieve this purpose were the: (1) hardware aspect, i.e., the complimentary use of ICT (confirming the whereabouts of children with a location information system, sharing this information on monitored children through an information sharing system, and providing emergency contact among residents with an information delivery system). The (2) software aspect involved the activities of residents (going on patrols, giving precautionary information at fixed points, and sending out information from FM stations). (3) Collaboration involved the strengthening of regional alliances (e.g., cooperation among organizations such as schools and the police, and activities by young support corps and women's groups to prevent crime.)

Interviews with Lay Subjects and Specialists

Comments by four stakeholders, as qualitative data, are given here. The interviews were carried out in August, 2008.

Mr. A (a representative of NPO: a lay subject)

> "Problems with living have occurred at sites where people live. First, only residents notice these. There are many things that only residents who live in the region understand. Furthermore, problems in our region need to be solved here, because our own children and grandchildren might be future victims of crime. The problems in our community, such as those regarding welfare and education, became evident during our activities to prevent crime. It is possible to solve these through cooperation and collaboration even though individual efforts are minimal. Continuing activities by using individual strengths is important. We, as

residents, and also school teachers, police officers, and ICT engineers, have had numerous discussions to analyze the situation in our community and methods of improving this."

Mr. B (a patrol member: a lay subject)

"I have been doing these activities for more than two years. I want to repay the kindness the region, where I grew up, has given me. We have continued patrols into the night as well as in the morning, and have found they have had a certain positive effect. I feel the patrols are beneficial. It is good that safety and security have been improved in my community. I am going to continue these activities and I want to enlist an even larger group of patrol members."

Mr. C (An ICT systems engineer: a specialist)

"I am a professional and specialist in systems development. However, residents know how the ICT system is used in the Tomioka district better than I do, i.e., who to use it with, where to use it at, and when and how to use it. Information literacy also varies according to residents. Although I developed the original specifications for the system, residents express their opinions according to conditions in actual use. I then give them feedback on their opinions, and the residents make additional comments. The system is customized in this way though intercommunication and we have obtained a system that is tailor-made for the Tomioka district."

Dr. D (A Professor majoring in crime prevention: a specialist)

"The main factor responsible for the success of these activities has been collaboration. Specialists do not one-sidedly offer knowledge, and residents do not continue to experience activities at random. Various subjects with certain strengths cooperated in these activities. Moreover, the devices for group participation operated really well. The effect crime-prevention activities have had in the Tomioka-district has steadily increased. The incidences of bag snatching and loitering have sharply declined. There are not that many regions where crime-prevention activities are continuing as they are here. The methods and techniques we have used can be used as references by other regions."

Dynamics of Interaction between LK and SK

Further collaboration among not only residents but also specialists such as engineers, police, schools, and crime-prevention researchers was found to have developed using the data obtained from the interviews and the observation of actual activities. Individual subjects in such collaboration send their knowledge through continuous intercommunication.

Figure 2 shows the flows of interactions between LK and SK in the crime-prevention activities in the city of Sakai. Specialists such as engineers and professors develop theories and methods within a scientific framework: however, the society or region moves as a whole. Lay subjects, on the other hand, have experiential perceptions or wisdom. It is the wisdom and experience of residents who live in the region that ties the whole together. Ideas and knowledge are re-designed according to movements within real societies or regions.

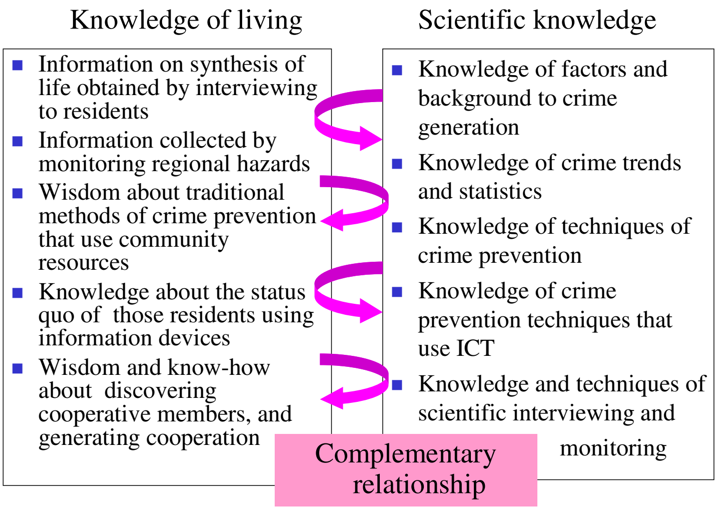

Knowledge of living
Scientific knowledge
Information on synthesis of life obtained by interviewing to residents
Information collected by monitoring regional hazards
Wisdom about traditional methods of crime prevention that use community resources
Knowledge about the status quo of those residents using information devices
Wisdom and know-how about discovering cooperative members, and generating cooperation
Knowledge of factors and background to crime generation
Knowledge of crime trends and statistics
Knowledge of techniques of crime prevention
Knowledge of crime prevention techniques that use ICT
Knowledge and techniques of scientific interviewing and monitoring
Complementary relationship

Fig. 2. Interaction between LK and SK in crime prevention activities in Sakai

There are four main factors that enabled communication between LK and SK in Sakai: (1) Lay subjects and specialists alike had a strong sense of consciousness in recognizing themselves as people concerned with the welfare of others. LK demonstrated its enormous influence in raising a strong sense of consciousness about people's concerns. It was important in communications between LK and SK for lay subjects to participate with a high degree of subjectivity. (2) An appropriate mechanism for further collaboration should be developed. (3) It is crucial to establish trust not only among lay subjects but also between lay subjects and specialists. The veracity of information, the honesty of the sender, and sharing of the same values are needed to generate trust. (4) Not depriving the region of potential problem-solving capabilities is important. Problem solving could not be sustained if specialists merely gave regions equipment or knowledge.

4 Conclusion and Future Work

This paper tried to clarify the uniqueness of LK and SK as well as the significance of communication between them, and to examine the potentiality of and problems with communication between the two kinds of knowledge as chance discovery with some empirical data. Specialists and lay subjects were found to have outstanding knowledge in mutually different contexts even though they had mutual limitations. A questionnaire revealed that although most lay subjects were proud of their own LK, they accepted SK as indispensable and helpful. The case study in the city of Sakai actually demonstrated that the quite real and complicated problem of crime prevention could be solved with the collaboration of lay subjects and specialists using their mutual knowledge on an equal footing.

The conditions under which lay subjects form their views and their methods of communication that depend on attributes such as their age, occupation, and family structure should be clarified in future work. Cross-cultural studies on communication between LK and SK also have to be conducted. Moreover, significance of "trust" in the double helical model should be examined. It can be argued that the deficit is not just a deficit in knowledge but more a deficit in trust that the necessary procedures are followed correctly.

Acknowledgements

This study was supported by the Japan Society for the Promotion of Science (JSPS) by a Grant-in-Aid for Scientific Research (B).

References

Arimoto, T.: Communication and Structural Transition of Science and Technology. Research for Science Education 31(4) (2007)

Chiba, K.: Science Communication: Five Techniques to Communicate Science. In: Nihonhyoronsya, C.G. (ed.) Local Knowledge: Further Essays in Interpretive Anthropology. Basic Books, New York (1983)

Hirakawa, H., Shirabe, M.: Live in High-technology Society. Kitakisyuppann (2003)

Horii, H.: Science and Technology for Society to Achieve Safety and Security, Tokyodaigakusyuppannkai (2006)

Inglis, J.T. (ed.): Traditional Ecological Knowledge: Concepts and Cases. Canadian Museum of Nature (1993), http://www.idrc.ca/en/ev-9321-201-1-DO_TOPIC.html

Kobayashi, T.: The Era of Trans-Science: Tie Science/Technology and Society. NTTsyuppan (2007)

Ohsawa, et al.: Data-Analysis Competition, Marketing Engineering Section. The Operations Research Society of Japan (2001)

Ohsawa, Y., McBurney, P. (eds.): Chance Discovery. Springer, Heidelberg (2003)

Wynne, B.: Public Understanding of Science. In: Jasanoff, S., et al. (eds.) Handbook of Science and Technology Studies. Sage, Thousand Oaks (1995)

Automatically Estimating and Updating Input-Output Tables

Ting Yu[*], Manfred Lenzen, Chris Dey, and Jeremy Badcock

Centre of Integrated Sustainability Analysis, Physics Building A28, University of Sydney,
NSW 2006, Australia
t.yu@physics.usyd.edu.au

Abstract. This paper presents an integrated intelligent system being capable of automatically estimating and updating large-size input-output tables. The system in this paper consists of a series of components with the purposes of data retrieval, data integration, data analysis, and quality checking. This unique system is able to interpret and follow users' XML-based query scripts, retrieve data from various sources and integrate them for the following data analysis components. The data analysis component is based on a unique modelling algorithm which constructs the matrix from the historical data and the spatial data simultaneously. This unique data analysis algorithm runs over the parallel computer to enable the system to estimate a large-size matrix. The result demonstrates the acceptable accuracy by comparing a part of the multipliers with the corresponding multipliers calculated by the matrix constructed by the surveys.

1 Introduction

In the theoretical economics, the input-output model of economics uses a matrix representation of a nation's (or a region's) economy to predict the effect of changes in one industry on others and by consumers, government, and foreign suppliers on the economy [1]. Because the economic constantly evolves, the input-output model needs to be updated at least annually to reflect the new circumstance. Unfortunately, in most countries such as Australia, the input-output model is only constructed every 3-4 years, because the large amount of monetary and human cost is involved. The Centre for Integrated Sustainability Analysis (ISA), University of Sydney, is developing an integrated intelligent system to estimate and update the input-output model at different level on a regular basis.

The input-output model often consists of a time series of matrices which may have temporal stability or temporal patterns. At the same time, within a given time period, extra information regarding certain parts of the matrix is often available from various government departments or other public or private organizations. However, most of this information is often incomplete and only gives a snapshot of a part of the underlying model. Apart from the massive data, hundreds of years of research has accumulated substantial amount of general knowledge of the national economic. Any

[*] Corresponding author.

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 42–49, 2009.

researcher could utilize this public knowledge to facilitate their discovery. On the contrast, other knowledge discovery activities often do not have such rich resource.

A time series of input-output models represents the evolution of industry structure within and between regions, where the region is defined as a geographic concept. It is a spatio-temporal knowledge discovery process with the help of rich domain knowledge. Including time introduces additional complexity to the geographic knowledge discovery [2]. This paper presents a novel algorithm which estimates and updates the economic matrix for the general equilibrium theory.

2 System Design

The whole system consists of functional components: data retrieval, data integration, data modelling and model presentation. The row data is retrieved from various data sources, and restructured and integrated into a data mining model. Then the data model is fed into the data modelling algorithm and consequently solved by the optimization engine. The result from the data modelling algorithm is the final result that is an estimated matrix.

The data retrieval component acts as interfaces to all types of datasets including macro and micro economic data that are stored in various formats such as Excel files, databases etc. The data integration component unifies these heterogeneous datasets to a single format, integrates and restructures the data retrieved by the previous component and presents the result for data mining. The data modelling component is the core of the whole system. In this component, a unique data modelling algorithm is designed to estimate the matrix.

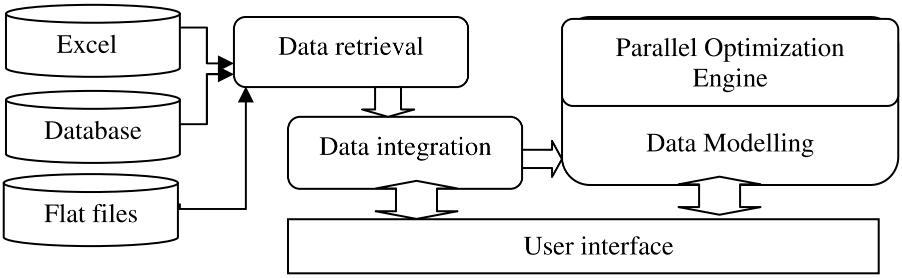


Fig. 1. System Architecture

2.1 Data Integration

The data integration component includes two main sub-modules: the structure builder and the model constructor (See Figure 2). Within the model constructor, there are two processes to restructure the data: 1) require the interfaces to retrieval data from various sources and integrate them, and 2) restructure and assign the meaning to the data according to the previous tree structure and users' specification and populate the mining model.

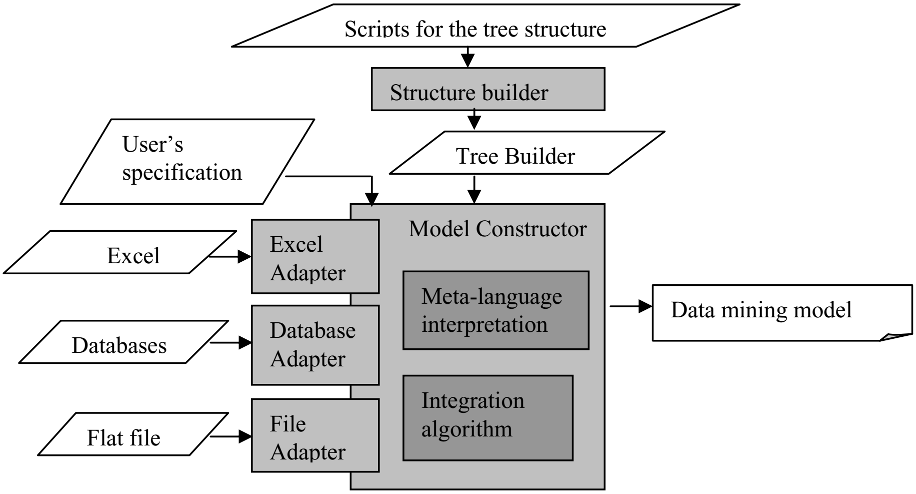


Fig. 2. Data Integration Component

			China (1)	
			Shoe (1)	Retail (2)
Australia (1)	NSW (1)	Sheep (1)	$X_1 = 0.23$	X_2
		Oil (2)	X_3	X_4
	VIC (2)	Sheep (1)	X_5	X_6
		Oil (2)	X_7	X_8

Fig. 3. An Example of the Matrix Defined by the 3-level Tree and the 2-level Tree

The first step is to construct the tree structure. The tree structure is pre-required for restructuring data collected from various sources. An example of the tree structure (See Fig 3) is a three-level tree representing the Australian Economic, one branch of which represents the sheep industry section within the New South Wales, a state of Australia. If the numerical indices are employed instead of their names, the sheep industry section within the New South Wales, a state of Australia can be written in [1,1,1] which means the first leaf in the first branch of the first tree.

The row and column of the matrix is defined by this tree structures, thereby the matrix is defined by the tree structures. The tree structure is unnecessarily with three levels. For example, a matrix (see Figure 3) can be organized by one three-level tree at the row side and one two-level tree at the column side. The coordinate of one entry, say X_1, can be defined as by [1,1,1] at the row side and [1,1] at the column side. That means the entry, X_1, defined by a three-level tree structure and a two-level tree structure at the column side. The tree structure is crucial to assign the meaning to the data retrieved from various sources, since the coordinates of entries are completely determined by it.

Considering the difference between applications, a dynamical structure of resultant matrix provides the flexibility to expand this software system to different application. On the other hand, the flexibility of the structure makes the system to be available to various level of implementation. For example, there is huge difference between the structures of resultant matrix at the national and at the corporate level, as the operations within a corporate are much simpler than those of a nation in the most cases. The dynamic of the structure is introduced by a multi-tree structure in Figure 3.

Considering the complexity of the model, a Meta language is introduced to provide users' an easy way to organize their data. The Meta language must be compact and accurate to make the description to be readable and useful. It is unrealistic to write hundred thousands of code to describe a single model at a daily base. The meta language we create is based on the coordinate of the valuable in the resultant matrix. For example, the coordinates of one entry is written as [1, 1, 1 -> 1, 1]. The value of this entry X_1 is indicated as the (0.23) [1, 1, 1 -> 1, 1] (See Figure 3). The system will fill the 0.23 in the cell with the coordinate [1, 1, 1] at the row side and [1, 1] at the column side. Consequently, this script indicates that 0.23m dollar worth of sheep products are transferred to the shoe industry in China.

2.2 Spatio-temporal Modelling with Conflict Information

The data analysis component is the core engine of the whole system. In this component, a unique modelling algorithm is designed to estimate the matrix. This modelling algorithm utilizes two types of information: the historical information which contains the temporal patterns between matrices of previous years, and the spatial information within the current year. For example, this spatial information can be the total output of the wool industry in Australia within the current year, or the total greenhouse emission of the car manufacture industry in Australia. The previous tree structure is employed here to represent the geographic concept hierarchies within the spatial information. The modelling algorithm can be written in the format of an optimization model as below:

$$Min[\frac{dis(X-\overline{X})}{\varepsilon_1}+\sum\frac{e_i^2}{\varepsilon_{i+1}}]\text{, Subject to: } G_1X+E=C_1 \qquad (1)$$

$$G_2X=C_2$$

$$X\geq 0$$

where X is the target matrix to be estimated , $\overline{X}$ is the matrix of the previous year, E is a vector of the error components $[e_1,...,e_i]^T$, *dis* is a distance metric which quantifies the difference between two matrices, e.g. $\sum(X_i-\overline{X}_i)^2$ in this case. G is the coefficient matrix for the local constraints, and C is the right-hand side value for the local constraints. The idea here is to minimize the difference between the target matrix and the matrix of the previous year, while the target matrix satisfies with the local regional information to some degree. For example, if the total export of the sheep industry from Australia to China is known as c_1, then $GX+E=C$ can be

$[1,1]\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} + e_1 = c_1$. The element e_i in E represents the difference between the real value and estimate value, for example, $e_1 = c_1 - [1,1]\begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$. The reason why it is introduced is to solve the *conflicting information*. Very often the data collected from different sources is inconsistent between each other, and even conflicting. Here e_i is introduced to balance the influence between the conflicting information, and reaches a tradeoff between the conflicting information. This modelling algorithm assumes the temporal stability, which assumes the industry structure of a certain region keeps constant or has very few changes within the given time period. By this assumption, this algorithm considers the matrix for the previous year as a reasonable proxy to the current year. Within two successive years, dramatic change of the industry structure is relatively rare, and this assumption has a good ground.

The reason why the spatio-temporal modelling algorithm is suitable to this system is due to the unique characteristics of the datasets that the system aims to process. The datasets often contain the temporal patterns between years, such as the trend of the production of certain industry sections, and also much spatial information regarding the total production within a certain region such as national total emission and state total emission. Even more, the datasets also contains the interrelationship between the industries within a given region or between regions. On the other hand, it is very common that either of datasets is not comprehensive and imperfect and even the conflicts between the datasets exist. Thereby, the modelling algorithm is required to consolidate the conflicted datasets to uncover underlying models, and at the same time, the modelling algorithm is required to incorporate the spatial information and keep the spatial relationship (such as dependency and heterogeneity [3]) within datasets.

2.3 Parallel Optimization

In real world practice, the previous modeling algorithm often processes matrix with dimensions over 1000-by-1000. In the foreseeable future, the size of estimated matrix will increase over 100,000-by-100,000. This requires the algorithm to have extremely outstanding capacity of processing large datasets. In order to address this problem, one parallel optimization algorithm is designed as the solver. The key idea is to divide the constraints into a few subsets of constraints, and then to do optimization against the subset of constraints respectively instead of the whole set of constraints. The simplest case is that the original optimization problem is rewritten as a set of sub-problems.

Sub-problem 1 (soft constraints):	Sub-problem 2 (hard constraints):	Sub-problem 3 (nonnegative constraints):
$Min[\frac{(X-\overline{X})^2}{\varepsilon_1} + \sum \frac{e_i^2}{\varepsilon_{i+1}}]$ Subject to: $G_1X + E = C_1$	$Min[\frac{(X-\overline{X})^2}{\varepsilon_1}]$ Subject to: $G_2X = C_2$	$Min[\frac{(X-\overline{X})^2}{\varepsilon_1}]$ Subject to: $X \geq 0$

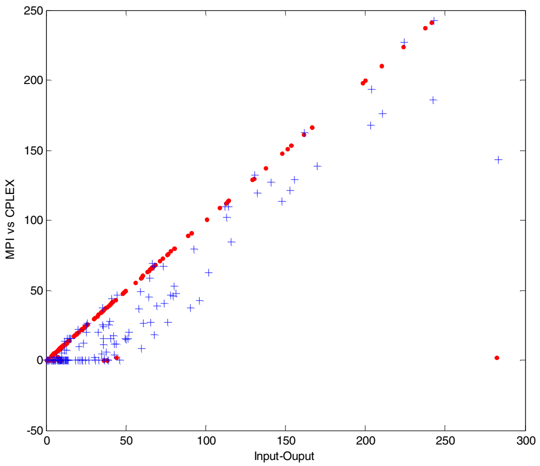


Fig. 4. Results from the CPLEX (Blue Cross Points) vs. results from the Parallel Optimization (Red Dot Points) by comparing with the real input-output data. The x-axis represents the real input-output table.

The results from the sub-problems are combined as a weighted sum which consequently acts as a start point for the next iteration. Suppose the result from the ith sub-problems is $P_i(X_n)$, the weighed sum is written as $X_{n+1} = X_n + L*[\sum w_i P_i(X_n) - X_n]$, where L is the relaxation parameter. This method is a special case of the parallel projection method [4]. Because the objective function of this particular problem is quadratic, thereby convex and the constraints are linear thereby convex as well, the optimization process is simpler than general projection methods. This parallel optimization algorithm is implemented over the Message Passing Interface (MPI). For the purpose of demonstration, the performance is compared with a commercial optimization package, the CPLEX by using the same test dataset concluding 12-by-12 entries with 57 constraints.

According to the experiments (see Figure 4), the parallel optimization estimates the underlying the matrix better as the linear relationship between its estimated result and the real matrix are very clear. The drawback is that the parallel optimization does not prevent the estimated value from becoming negative. The first sub-problem requires the estimation to be positive or zero. However in Figure 4, some estimated values are very small negative.

3 Experimental Results

The direct evaluation of a large-size matrix is a rather difficult task. A thousand-by-thousand matrix contains up to ten million numbers. Simple measurements such as the sum do not make too much sense, as the important deviation is submerged by the total deviation which normally is far larger than the individual ones. The key criterion here is the distribution or the interrelationship between the entries of the matrix: whether

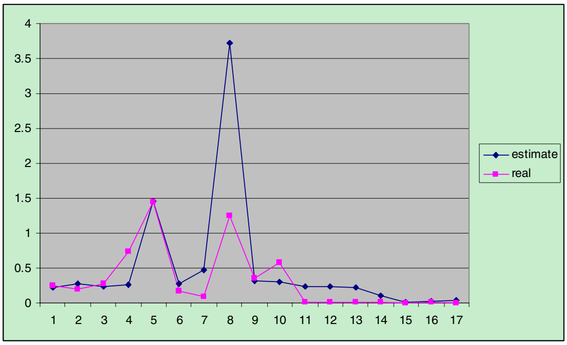


Fig. 5. Comparison between two series of multipliers

the matrix reflects the true underlying structure, not necessary the exactly right value, at least the right ratios.

The multipliers in the input-output framework reflect the aggregated impacts of the final demand changes on the upstream industries [1]. The information contained by the multipliers is very similar to the sensitivity analysis in general statistics. The general formula of constructing the multipliers is: $M = D(I - A)^{-1}$ where M is the multiplier, I is the identity matrix, D is the change in the final demand, and A is the matrix, each entry of which is $X_i / \sum_{i=1}^{n} X_i$. Here, X_i is a value from the matrix estimated by the equation (1).

This sensitivity multiplier counts the impact of any change of outputs on the whole upstream inputs, and not only the direct inputs. Any deviation occurring in the upstream inputs from the underlying true structure will be amplified and reflected on the multipliers. Thereby, the multipliers send an indirect warning signal to imply the structural deviation occurring on the upstream inputs.

As a case study, a matrix aims to calculate the total water usage of the different industries in Australia. A part of the data is collected from the Water Account reports produced by the Australian Bureau of Statistics [5]. The full Australian economy consists of 8 states with 344 industry sections plus7 final demands and 6 value added sections. Totally, the Australian input-output table is a 2808-by-2800 matrix, containing 7,862,400 entries. In order to estimate the matrix, more than 260,000 constraints are included.

From the below plot (see Figure 5) comparing the two series of the multipliers, two series basically follow the same trends, which indicate the industry structure is estimated properly. However the estimated multipliers are more volatile than the true underlying multipliers. This phenomenon indicates the estimated multipliers amplify the errors introduced to the upstream industries. Another reason of the difference is the underlying structure change within a given industry. In Figure 5, the big gap between two series at 8 indeed indicates from 1999 to 2004, the Australian rice industry dramatically reduces its rice production due to the continuous draught, but imports more and more rice from other nations. As the price has been inflated and the water usage is dropping, the ratio of the water usage by price is dropping.

4 Conclusion

This system is an integrated data analysis system for updating a large-scale matrix. The unique characteristics of the data determine the data analysis system must be capable of dealing the temporal and spatial data simultaneously. At the same time, the large size of the estimated matrix requires the system to process a large amount of data efficiently. This paper presents a completed data analysis system starting from data collection to data analysis and quality checking. According to the result of the experiments, the system successfully produces the matrix, and makes it a rather easy task without a huge amount of work to collect and update both data and model. Before this system, this kind of collection and updating work costs months of work, but now it takes only a few days with the consistent quality.

As the temporal stability is a major assumption, the further developments will emphasis on how to deal with the major structure changes. This research can be applied to broader horizon such as the Markov transient matrix.

References

1. Miller, R.E., Blair, P.D.: Input-output Analysis, Foundations and Extensions. Prentice-Hall Inc., Englewood Cliffs (1985)
2. Miller, H.J., Han, J.: Geographic Data Mining and Knowledge Discovery. CRC, Boca Raton (2001)
3. Miller, H.J.: Geographic Data Mining and Knowledge Discovery. In: Wilson, J., Fotheringham, A.S. (eds.) The Handbook of Geographic Information Science, Wiley-Blackwell (2007)
4. Combettes, P.L.: A Block-iterative Surrogate Constraint Splitting Method for Quadratic Signal Recovery. IEEE Transactions on Signal Processing 51(7), 1771–1782 (2003)
5. 4610.0 - Water Account, Australia. 2004-05, The Australian Bureau of Statistics: Canberra

Context-Aware User and Service Profiling by Means of Generalized Association Rules*

Elena Baralis[1], Luca Cagliero[1], Tania Cerquitelli[1], Paolo Garza[1],
and Marco Marchetti[2]

[1] Politecnico di Torino - Dipartimento di Automatica e Informatica - Torino, Italy
{elena.baralis,luca.cagliero,tania.cerquitelli,paolo.garza}@polito.it
[2] Telecom Italia Lab - Torino, Italy
marco1.marchetti@telecomitalia.it

Abstract. Context-aware applications allow service providers to adapt their services to actual user needs, by offering them personalized services depending on their current application context. Hence, service providers are usually interested in profiling users both to increase client satisfaction, and to broaden the set of offered services.

Since association rule extraction allows the identification of hidden correlations among data, its application in context-aware platforms is very attractive. However, traditional association rule extraction, driven by support and confidence constraints, may entail either (i) generating an unmanageable number of rules in case of low support thresholds, or (ii) discarding rare (infrequent) rules, even if their hidden knowledge might be relevant to the service provider. Novel approaches are needed to effectively manage different data granularities during the mining activity.

This paper presents the CAS-Mine framework to efficiently discover relevant relationships between user context data and currently asked services for both user and service profiling. CAS-Mine exploits a novel and efficient algorithm to extract generalized association rules. Support driven opportunistic aggregation is exploited to exclusively generalize infrequent rules. User-provided taxonomies on different attributes (e.g., a geographic hierarchy on spatial coordinates, a temporal hierarchy, a classification of provided services), drive the rule generalization process that prevents discarding relevant but infrequent knowledge.

Experiments performed on both real and synthetic datasets show the effectiveness and the efficiency of the proposed framework in mining different types of correlations between user habits and provided services.

Keywords: Generalized association rules, knowledge discovery, context-aware data.

1 Introduction

Context-aware systems acquire and exploit information on the user context to tailor services to the particular user, place, time, and/or event. Research activities

* This work was supported by a grant from Telecom Italia Lab.

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 50–57, 2009.

on context-aware computing have been devoted both to exploring the different dimensions of context-awareness [8], and to implementing different context-aware applications (e.g., in the medical domain [15], for mobile phones [6]). Context could consist of any circumstantial factors or application context users are involved in. Thus, context-awareness means that the system is able to exploit context information. A system is context-aware if it can extract, interpret and use context information and adapt its functionalities to the current usage context [9]. Provided services could be personalized by exploiting either the current context of the user [4] or historic context and behavior of the user [5]. An in-depth literature review on context-aware systems has been presented in [8].

Authors in [16] first proposed to exploit statistical machine learning techniques (e.g., rule induction, neural networks, Bayesian networks) to build predictive user models. These models have been exploited in different context-aware settings (e.g., a smart home [12], or a smart office [11]). Application and user profiling, instead, have been addressed in [10]. The proposed solution exploits rule based and naive Bayesian classifiers. In particular, different service and application models are tailored to the user and to the situation in which she is involved. These models are then exploited to suggest applications and services on what the user might interest in her current situation.

Association rule extraction [1] is a widely used exploratory technique allowing the discovery of hidden correlations among data. Its application in context-aware platforms to profile both users and services is very attractive. However, traditional association rule algorithms are not effective in mining context datasets because of the high detail level of the information (e.g., GPS coordinates). When low support thresholds are enforced, an unmanageable number of rules is extracted, while high support thresholds discard rare rules even if their knowledge might be relevant for the service provider. To address this issue, rules should be extracted at a higher abstraction level (i.e., generalized).

This paper presents the CAS-Mine framework to efficiently discover relevant relationships between user context data and currently requested services. Extracted rules may be exploited both for user and service profiling. To this aim, CAS-Mine exploits a novel and efficient algorithm, called GenIO, to extract generalized association rules. The GenIO algorithm extends the concept of multi-level rules [7] by performing an opportunistic extraction of generalized itemsets. It exploits (user provided) taxonomies to drive the itemset generalization process and efficiently extract generalized itemsets. Instead of extracting itemsets for all levels of the taxonomy and post-pruning them [7], the generalization step over the taxonomy is support driven, i.e., it generalizes an item climbing up the taxonomy if and only if its support is below the support threshold. The generalization process stops when the newly generalized item is above the support threshold.

Preliminary experiments performed on both real and synthetic datasets show the effectiveness and efficiency of the CAS-Mine framework in highlighting interesting rules to characterize users and services.

The paper is organized as follows. Section 2 motivates our work. Section 3 presents an overview of the CAS-Mine framework and describes its main features. In Section 4 preliminary experiments to validate the proposed framework are reported, while Section 5 draws conclusions and discusses future work.

2 Motivations

A structured context dataset holds information on service requests performed by users and the corresponding application context. Each data element is a set of items describing a service request and its context. Each item is a couple *(AttributeName, Value)*. The *AttributeName* describes the represented information (e.g., *user identifier*, *service*, *time*), while *Value* is the actual value of the corresponding attribute (e.g., *ID54*, *weather*, *4:06pm*). A generalized item is defined by means of a user-defined hierarchy of aggregation (i.e., taxonomy) over values in the attribute domain. For example, the *position* attribute may describe GPS coordinates (e.g., 45.438:12.335). The couple *(position, 45.438:12.335)* is an item (at the lowest level in the hierachy), while the couple *(position, office)* is a generalized item. Thus, *office* aggregates all GPS coordinates related to the office physical location.

Generalized association rules are represented in the form $A \Rightarrow B$ (s%,c%), where A and B are sets of (possibly generalized) items, and s% and c% represent support and confidence. The support is the prior probability of A and B (i.e., its observed frequency in the dataset). The confidence is the conditional probability of B given A and characterizes the "strength" of a rule. User activity may be characterized by the following association rule

$$\texttt{user}\text{: John, }\texttt{time}\text{: 6.05 p.m.} \Rightarrow \texttt{service}\text{: Weather } (s = 0.005\%, c = 98\%)$$

This specific rule is characterized by a very low support and is not extracted, because the extraction process would become unfeasible. By generalizing the time attibute on a time period, and the user on a user category, the following generalized rule may be obtained.

$$\texttt{user}\text{: employee, }\texttt{time}\text{: 6 p.m. to 7 p.m.} \Rightarrow \texttt{service}\text{: Weather}$$
$$(s = 0.2\%, c = 75\%)$$

If the obtained rule is still below the support threshold, the generalization process performs a further aggregation step on the time hierarchy.

$$\texttt{user}\text{: employee, }\texttt{time}\text{: Evening} \Rightarrow \texttt{service}\text{: Weather } (s = 1.5\%, c = 65\%)$$

The generalization process greedily continues until the obtained rule is above the selected support threshold. Thus, generalization allows highlighting interesting correlations which would be lost because of their low support at the lowest level of the hierachy.

3 The CAS-Mine Framework

CAS-Mine is a framework to efficiently profile both users and services. Thus, it allows shaping service supply by considering the context to which the user belongs. By discovering recurrent patterns involving user habits and requested services, providers may partition users into a set of well-known categories for which supplied services may be modeled and personalized.

CAS-Mine exploits the GenIO algorithm, a novel and efficient algorithm to discover generalized association rules. The mining process is driven by a set of (user-provided) taxonomies which allow the conceptual aggregation of items in more abstract categories. The main blocks of the CAS-Mine framework are reported in Figure 1, while a more detailed description of each block is presented in the following.

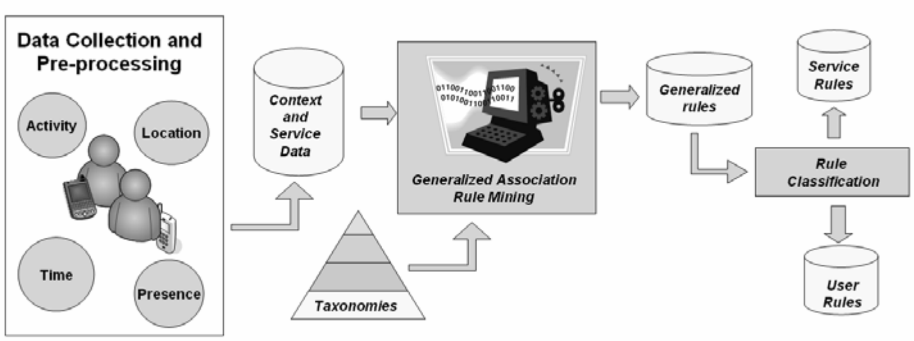


Fig. 1. The CAS-Mine Framework Architecture

3.1 Data Collection and Pre-processing

The data collection and pre-processing block manages data collection from a large number of mobile devices which provide information on the user context and on the supplied services (e.g., temporal information, GPS coordinates, service description).

The data collection block receives in input the raw context data provided by different, possibly heterogeneous, sources and integrates them into a common data structure. During this process, irrelevant and redundant information is also removed. Cleaned and integrated data are finally stored into a common repository.

3.2 Generalized Association Rule Mining

This block performs the extraction of generalized association rules. Extraction is performed in two steps: (i) frequent generalized itemset extraction and (ii) rule generation from the extracted frequent itemsets. Since itemset mining is the most computationally intensive step [1], the novel contribution of the GenIO algorithm focuses on itemset mining. The second step exploits Goethal's Rules software [3], possibly enforcing confidence constraint.

Given a dataset, a set of user-provided multi-level taxonomies (at most one for each attribute) and a minimum support threshold, the GenIO algorithm extracts all the frequent not-generalized itemsets and the set of generalized itemsets which represent the generalization of the knowledge associated to infrequent not-generalized itemsets.

GenIO performs a support-driven opportunistic aggregation of infrequent itemsets, thus avoiding exhaustive multi-level extraction followed by post-pruning. Hence, the GenIO algorithm reduces the cardinality of mined itemsets with respect to well-known traditional multi-level algorithms [7]. GenIO successfully tackles both excessive pruning and computationally hard exhaustive multi-level extraction, thus providing a good trade-off between itemset specialization and aggregation.

The itemset mining process, driven by the generalization step, is shortly described in the following. A detailed description of the GenIO algorithm is provided in [2]. GenIO is an Apriori-like [1] extraction algorithm, which performs level-wise itemset mining. In general, Apriori-like algorithms, at a generic level i, extract all the frequent itemsets of length i. Candidate itemsets of level $i + 1$ are generated by combining all the frequent itemsets of length i. Frequent itemsets of level $i + 1$ are then obtained by enforcing the support constraint.

The GenIO algorithm exploits rule generalization to extract knowledge on infrequent, but possibly interesting, itemsets. At each extraction level, before pruning itemsets not satisfying the support threshold, generalizes them by climbing up the generalization hierachy of the corresponding items. A generalized entry is a conceptual aggregation of different items at a lower level (e.g., *Communication service* represents lower level items *SMS service* and *CALL service*), thus it is more likely to be frequent. Taxonomies are exploited to drive the generalization process (see Section 2). GenIO climbs up each taxonomy in a stepwise fashion, until either support constraint is satisfied, or the highest generalization level is reached.

3.3 Rule Classification

The rule classification block categorizes generated rules in classes to effectively exploit them for different context-aware profiling. Since service providers are tipically interested in profiling both users and services, the CAS-Mine framework currently identifies two classes of association rules: (i) User rules and (ii) service rules. User rules characterize user habits at any aggregation level. These rules allow service providers to offer personalized services tailored to the current context of the user. Hence, provided services can be adapted to actual user needs. Service rules describe service characteristics, at any hierarchical level, without specific user information. These rules allow service providers to adapt service provisioning to the current context, independently of the requesting user (e.g., by providing a different bandwidth in different time periods).

4 Preliminary Experimental Results

We evaluated the CAS-Mine framework by analyzing (i) the characteristics and interestingness of extracted patterns on a real dataset, and (ii) the scalability, in terms of execution time, of the proposed approach on a synthetic dataset. All experiments were performed on a 3.2-GHz Pentium IV system with 2 GB RAM, running Ubuntu 8.04.

4.1 Characterization of Extracted Rules

The real dataset, denoted as *mDesktop*, has been provided by Telecom Italia Lab (Tilab). The trial version of the Tilab mobile desktop application provides different services to users (e.g., weather forecast) on mobile devices. The mDesktop dataset is characterized by 4487 records with information on each requested service and the context of the requesting user (e.g., time, location). The dataset is characterized by the following taxonomies:

- date → month → year
- timestamp → hour → day period (AM/PM)
- service → class of service
- latitude:longitude → city → country
- phone number → call type (PERSONAL/BUSINESS)

Figure 2 reports statistics about the mining activity performed on the mDesktop dataset. By setting high minimum support thresholds (e.g., 10%), only very frequent rules are extracted. In particular, only the generalized rules composed by the top levels of the taxonomies are extracted (e.g., `location`: ITALY ⇒ `date`: 2008). These rules are usually too general to provide interesting knowledge. Differently, when lower threholds are enforced, the extracted generalized rules include non-top level elements of the taxomonies, which may provide more actionable knowledge.

In the following, we analyze two different rule subsets, which show the effectiveness of the GenIO algorithm in supporting both user and service profiling.

User profiling rules. These rules deal with context-aware profiling for specific users whose habits show some kind of recurrence. In particular, for an arbitrary user, the GenIO algorithm highlights the service type the user is mainly interested in, the context in which requests are commonly submitted, and the service parameters. The following two rules (support threshold=1%, absolute threshold=45) discover valuable knowledge about an anonymous client, denoted here as *Rossi*:

A) `user`: Rossi ⇒ `service`: CALL ($s = 1.27\%, c = 53\%$)
B) `user`: Rossi ⇒ `service`: SMS ($s = 1.14\%, c = 47\%$)

The above rules highlight that user *Rossi* is interested in two specific services, CALL and SMS, with confidence close to 50%. Thus, they provide a relevant knowledge on this user preferences. When the (higher) support threshold 2% is enforced (absolute threshold = 90), the following generalized rules are extracted.

A) `user`: Rossi ⇒ `hour`: PM ($s = 2.25\%, c = 94\%$)
B) `user`: Rossi ⇒ `service`: Communication ($s = 2.41\%, c = 100\%$)

The first rule shows an intensive system usage for user *Rossi*, especially during the afternoon/evening (confidence 94%). Furthermore, the second rule, with confidence 100%, shows that *Rossi* is exclusively interested in the *Communication* service class, which contains the CALL and SMS services. A traditional mining

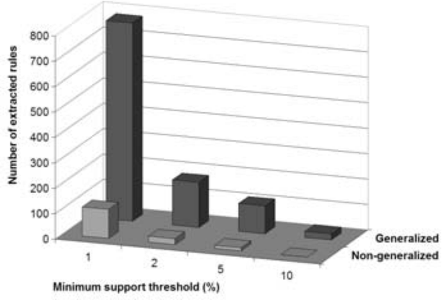


Fig. 2. Rule statistics

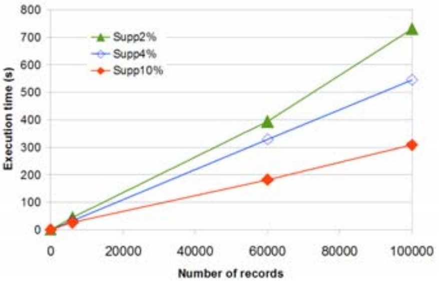


Fig. 3. Scalability on TPC-H datasets

algorithm would lose the above knowledge by pruning the lower level infrequent itemsets. GenIO capability of automatically climbing up the service taxonomy allows the extraction of these interesting high level rules.

Service profiling rules. These rules characterize frequently used services, independently of the specific user requesting them. The following generalized rules are extracted with support threshold 1% (absolute threshold = 45).
A) `date`: August, `hour`: PM ⇒ `service`: HOME ($s = 7.09\%, c = 94\%$)
B) `service`: CALL ⇒ `inout`: OUT ($s = 1.14\%, c = 89\%$)
The first rule shows that 94% of service requests submitted in the evening during the month of August are related to the HOME service. This knowledge may be exploited both to size system resources, hence providing a more efficient HOME service, and to select the first service to suggest to connected users. The second rule highlights the correlation between the service type and its parameters. In this case, call services are mainly exploited to perform outgoing calls.

4.2 Scalability

We analyzed the scalability of the GenIO algorithm with respect to the cardinality of transactions on syntethic datasets generated by means of the TPC-H generator [13]. By varying the scale factor parameter, tables with different cardinalities are generated. We generated datasets of size ranging from 6,000 to 100,000 transactions with 8 categorical attributes. We mined generalized itemsets from the *lineitem* table and we exploited the *part*, *nation*, and *region* tables to define taxonomies on line items.

Figure 3, which plots the extraction time for various supports, shows that the proposed algorithm scales well also for large datasets. Since the number of extracted itemsets grows for low supports (e.g., 2%), the process becomes computationally more expensive. However, the overall CPU time is still low, less than 730s for the lowest considered support and largest dataset.

5 Conclusion and Future Works

Context-aware applications exploit implicit context information (e.g., environmental conditions, location, time) to enhance the fulfillment of explicit user requests by providing personalized services tailored on the current application context of the user. In this paper we presented the CAS-Mine framework to support context-aware user and service profiling. CAS-Mine exploits a novel algorithm to efficiently mine generalized association rules. The mining process is driven by user-provided taxonomies on different attributes, which prevent discarding relevant but infrequent knowledge.

Future extensions of the framework will address (i) the automatic inference of taxonomies from the input context dataset, (ii) the exploitation of multiple taxonomies over a single attribute, and (iii) the application of the opportunistic generalization approach to more efficient rule extraction algorithms (e.g., LCM [14]).

References

1. Agrawal, R., Srikant, R.: Fast Algorithms for Mining Association Rules in Large Databases. In: Proceedings of the 20th VLDB conference, pp. 487–499 (1994)
2. Baralis, E., Cerquitelli, T., D'Elia, V.: Generalized itemset discovery by means of opportunistic aggregation. Technical report, Politecnico di Torino (2009), https://dbdmg.polito.it/twiki/bin/view/Public/NetworkTrafficAnalysis
3. Goethals, B.: Frequent Pattern Mining Implementations, http://www.adrem.ua.ac.be/~goethals/software
4. Bradley, N.A., Dunlop, M.D.: Toward a multidisciplinary model of context to support context-aware computing. Hum.-Comput. Interact. 20(4), 403–446 (2005)
5. Byun, H., Cheverst, K.: Utilizing context history to provide dynamic adaptations. Applied Artificial Intelligence 18(6), 533–548 (2004)
6. Hakkila, J., Mantyjarvi, J.: Collaboration in context-aware mobile phone applications. In: Hawaii International Conference on System Sciences, vol. 1, p. 33 (2005)
7. Han, J., Fu, Y.: Mining multiple-level association rules in large databases. IEEE Trans. Knowl. Data Eng. 11(5), 798–804 (1999)
8. Hong, J., Suh, E., Kim, S.: Context-aware systems: A literature review and classification. In: Expert Systems with Applications (November 2008)
9. Khedo, K.K.: Context-aware systems for mobile and ubiquitous networks. In: ICNICONSMCL 2006, p. 123 (2006)
10. Nurmi, P., Salden, A., Lau, S.L., Suomela, J., Sutterer, M., Millerat, J., Martin, M., Lagerspetz, E., Poortinga, R.: A system for context-dependent user modeling. In: Meersman, R., Tari, Z., Herrero, P. (eds.) OTM 2006 Workshops. LNCS, vol. 4278, pp. 1894–1903. Springer, Heidelberg (2006)
11. Oliver, N., Garg, A., Horvitz, E.: Layered representations for learning and inferring office activity from multiple sensory channels. Comput. Vis. Image Underst. 96(2), 163–180 (2004)
12. Tapia, E.M., Intille, S.S., Larson, K.: Activity recognition in the home using simple and ubiquitous sensors. In: Ferscha, A., Mattern, F. (eds.) PERVASIVE 2004. LNCS, vol. 3001, pp. 158–175. Springer, Heidelberg (2004)
13. TPC-H. The TPC benchmark H. Transaction Processing Performance Council (2009), http://www.tpc.org/tpch/default.asp
14. Uno, T., Kiyomi, M., Arimura, H.: LCM ver. 2: Efficient mining algorithms for frequent/closed/maximal itemsets. In: FIMI (2004)
15. Vajirkar, P., Singh, S., Lee, Y.: Context-aware data mining framework for wireless medical application. In: Mařík, V., Štěpánková, O., Retschitzegger, W. (eds.) DEXA 2003. LNCS, vol. 2736, pp. 381–391. Springer, Heidelberg (2003)
16. Zukerman, I., Albrecht, D.W.: Predictive statistical models for user modeling. User Modeling and User-Adapted Interaction 11(1-2), 5–18 (2001)

An ETL Tool Based on Semantic Analysis of Schemata and Instances*

Sonia Bergamaschi[1], Francesco Guerra[2], Mirko Orsini[1], Claudio Sartori[3], and Maurizio Vincini[1]

[1] DII-Università di Modena e Reggio Emilia, Italy
firstname.lastname@unimore.it
[2] DEA-Università di Modena e Reggio Emilia, Italy
firstname.lastname@unimore.it
[3] DEIS - Università di Bologna,Italy
claudio.sartori@unibo.it

Abstract. In this paper we propose a system supporting the semi-automatic definition of inter-attribute mappings and transformation functions used as an ETL tool in a data warehouse project. The tool supports both schema level analysis, exploited for the mapping definitions amongst the data sources and the data warehouse, and instance level operations, exploited for defining transformation functions that integrate data coming from multiple sources in a common representation. Our proposal couples and extends the functionalities of two previously developed systems: the MOMIS integration system and the *RELEVANT* data analysis system.

1 Introduction

Enterprise Information Systems provide a technology platform that enables organizations to integrate and coordinate their business processes[1]. The data warehouse represents definitely one of the most important components of information systems, since it enables business intelligence analysis on data coming from multiple sources. Traditional architectures of data warehouse systems rely on *extraction, transformation and loading (ETL)* tools for building and populating the data warehouse. Such tools support (a) the identification of relevant information at the source side, (b) the extraction of this information, (c) the customization and integration of the information coming from multiple sources into a common format, (d) the cleaning of the resulting data set on the basis of database and business rules, and (e) the propagation of the data to the data warehouse and/or data marts [10].

ETL processes are crucial for the data warehouse consistency, and are typically based on constraints and requirements expressed in natural language in the form of comments and documentations. Consequently, the aforementioned tasks are manually performed

* This work was partially supported by MIUR co-funded project NeP4B (http://www.dbgroup.unimo.it/nep4b) and LISEA Laboratory (Italy, Emilia-Romagna Region -PRRIITT, measure 3.4).

[1] From Wikipedia, http://en.wikipedia.org/wiki/Enterprise_Information_System

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 58–65, 2009.

by the designer or the administrator of the data warehouse. In the context of traditional databases, this fact does not represent a real big issue: 1) the processes requiring the manual user intervention involve data source schemata that are generally fixed. Thus, designers are asked to set up the ETL processes once per data source (during the start up phase); 2) all the data sources collected in the data warehouses typically belong to the same company and are known and managed by the same designers. Thus, technical documentations and personal knowledge are available and may be usefully exploited for building the data warehouse.

Nowadays, the actual business needs require enterprise information systems to have a great flexibility concerning the allowed business analysis and the treated data. Temporary alliances of enterprises, market analysis processes, the data availability on Internet push enterprises to quickly integrate unexpected data sources for their activities. Therefore, the reference scenario for data warehouse systems extremely changes, since data sources populating the data warehouse may not be directly known and managed by the designers. We have classified four major critical activities:

- **Automating Extraction processes.** Designers may no longer rely on internal documents, comments and previous knowledge on the data source contents. Moreover, the manual exploration of the data source contents may be really time-consuming. Techniques for identifying the information held by the data sources and extracting the relevant data for populating the data warehouse are required.
- **Automating Transformation processes.** Data from different sources may not be homogeneous, i.e. different metrics may be used for expressing similar values, synonyms may be used for representing the same values (and vice-versa the same value in different sources may describe a different concept), and values may be expressed with different granularity levels. Therefore, the data transformation processes are crucial for making attribute values from different data sources *comparable*: techniques to automatically perform data transformations are needed.
- **Relaxing the Transformation processes.** Deeply homogenizing transformations risk to flatten data warehouse structure and contents, thus allowing less accurate data analysis on such data. A balance should be pursued between ability of introducing new sources and preservation of structural and contents homogeneity allowing accurate business intelligence processes.
- **Speeding-up and easying ETL process execution.** Fast and simple ETL execution is a crucial competition factor when the sources populating the data warehouse are dynamic. Moreover, ETL tools should be able to manage different kinds of data sources, ranging from DBMS, to flat file, XML documents and spreadsheets.

In this paper we propose a tool for automating the extraction process and the formulation of transformation functions which provides a real support in dealing with the above issues. By means of a semantic analysis coupled with clustering techniques, the tool performs two tasks: 1) it works at the schema level, identifying the parts of the schemata of the data sources which are related to the data warehouse, thus supporting the extraction process; 2) it works at the instance level, grouping the attribute values semantically related thus defining a transformation function for populating with "homogeneous" values the data warehouse.

The work relies on the union and extension of two previously developed systems: the MOMIS integration system, that will be exploited for the semantic analysis of the source schemata and the semantic clustering techniques of the RELEVANT system that will be exploited for recognizing related schema elements and similar attribute values (see [4,6] and related work section).

The paper is organized as follows: next section introduces some related work and an overview of both the previously developed MOMIS and *RELEVANT* systems. Section 3 describes the new tool along with a running example on a real application scenario; in Section 4, some conclusion and future work are sketched out.

2 Related Work

Many approaches have been proposed for modeling ETL processes both at the logical and conceptual level. Several methods propose extensions of UML in order to describe ETL processes [8]. The advantage of those methods relies on the fact that they are based on a well-accepted, standard modeling language. Other approaches are based on 'ad hoc' techniques having the benefit of representing ETL processes without any restriction imposed by a previously defined, generic language [10].

Recently, techniques exploiting ontologies to manage the heterogeneity of the data in the mapping and transformation processes [9] have been proposed . As our approach, they follow an hybrid approach, where each schema is represented by its own ontology and a common shared vocabulary is provided, while the annotation process is manual and the vocabulary has been done by the designer.

Apart from research efforts, currently, there is a variety of ETL tools available in the market (see [7] for a survey). All major database vendors provide ETL solutions bundling them with their DBMS [1,3,2].

Our proposal relies on two previously developed systems: MOMIS and *RELEVANT*.

The **M**ediator Envir**O**nment for **M**ultiple **I**nformation **S**ources (MOMIS)[2] [4]) is a semiautomatic data integration system that follows a global-as-view approach to provide an intelligent access to heterogeneous, structured and semi-structured information sources. MOMIS is based on clustering techniques applied to a set of metadata collected in a Common Thesaurus in the form of relationships describing inter- and intra-schema knowledge about classes and attributes of the local source schemata. The relationships are: 1) extracted from descriptions of local schemata; 2) obtained from the relationships existing in the WordNet database between the meanings associated to the source elements; 3) inferred by means of Description Logics techniques.

RELEVANT [6] is based on the idea that an attribute domain includes values which may be clustered because *strongly related*. Providing a name to these clusters, i.e. the relevant values, we refer to a relevant value name which encompasses a set of values. More formally, given a class C and one of its attributes At, a **relevant value** for it, rv^{At} is a pair $rv^{At} = \langle rvn^{At}, values^{At} \rangle$. rvn^{At} is the name of the relevant value set, while $values^{At}$ is the set of values referring to it. For computing the relevant values, the user may combine three different similarity measures (namely (1) *syntactic*, mapping all the words of the attribute values in an abstract space, and defining a syntactic similarity

[2] See http://www.dbgroup.unimo.it/Momis/ for more details and publications.

function in such space; (2) *dominance*, introducing a sort of generalization relationship between values; (3) *lexical*, which identifies semantically related values expressed with a different terminology) and select between two clustering algorithms (a hierarchical and an overlapping algorithm) .

3 Making Semantic ETL with MOMIS and *RELEVANT*

The data integration methodologies implemented in the MOMIS and *RELEVANT* systems are extended for supporting the automation of ETL processes. Section 3.1 describes how the coupling of the two systems is exploited for implementing the extraction process, and in section 3.2 we focus on supporting the semi-automatic definition of a new transformation function for the data warehouse population.

3.1 Semantic Extraction from Data Sources

Our approach takes as input the data warehouse and the data source schemata (more than one source may potentially be added to the data warehouse with a unique extraction process) and computes the mappings between those schemata. Figure 1 shows the functional architecture of the proposed tool. We identify three main phases: firstly, by exploiting the methodology implemented in MOMIS the descriptions of the sources are extracted and a thesaurus of relationships between the schema elements is computed. Schema descriptions and the thesaurus are then exploited by an extension of *RELEVANT* for computing clusters of similar elements. Finally, by means of a cluster analysis tool, mappings between the data warehouse schema and the new data sources are defined. For each phase, a software component has been built with the following features:

1. The schema descriptions of data sources are extracted by means of wrappers. A wrapper-based technology allows the tool to deal with several kinds of data sources such as spreadsheets, xml documents, text files and databases.

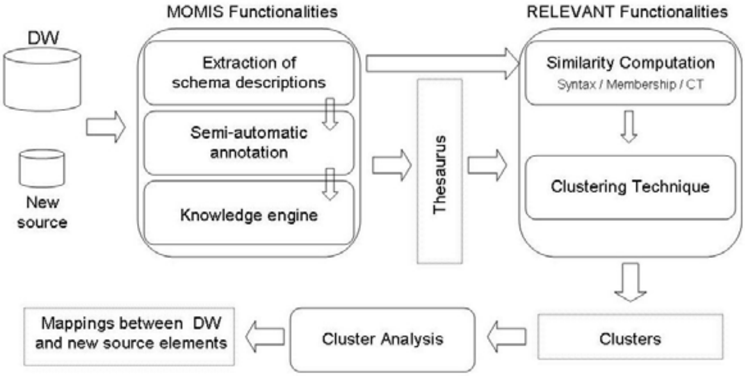


Fig. 1. Functional architecture of the semantic extraction tool

2. The annotation according to a reference ontology / database (in our tool, the WordNet[3] lexical database) allows the specification of a unique meaning to the schema description. In this way, it is possible to exploit the relationships among the referenced concepts to infer relationships between the annotated elements. We support this process by means of automatic techniques (see [5] for more details).
3. The knowledge engine is the component in charge of creating a thesaurus of relationships between the schema elements (in terms of classes and attributes). Three types of relationships are generated: relationship connecting synonym (SYN), broader terms/narrower terms (BT/NT) and generic relationships (RT). The process for extracting relationships is borrowed from MOMIS and is based on structural analysis (exploiting primary and foreign keys, attribute memberships), annotations and Description Logics techniques.
4. *RELEVANT* is applied to the descriptions of sources extracted in the first step with the aim of computing clusters of related attributes. *RELEVANT* has been extended by adding new similarity measures exploiting: 1) syntactic similarity, which compares the alphabets used for describing the attribute values; 2) memberships, which represents the closeness of attributes belonging to the same table; 3) semantic similarity, which takes into account the thesaurus of relationships between classes and attributes. Each similarity measure is represented by an affinity matrix, where a similarity value is computed for each attribute with respect to all other attributes according to the selected semantics. The combination of the values of each affinity matrix is parametric, thus allowing a setup phase where the user assigns an importance degree to some specific similarities. The application of a clustering algorithm (with a user-select threshold) generates clusters of similar elements.
5. Mappings are automatically generated by analyzing the clustering result. The following cases are possible:
   (a) A cluster contains attributes from the data warehouse schema and the new data sources: for each data warehouse attribute a mapping to each attribute in the cluster is generated.
   (b) A cluster contains only attributes of the new data sources: it is not exploited for the mapping generation. This cluster is due to the choice of a too selective clustering threshold.
   (c) A cluster contains only attributes of the data warehouse schema: it is not exploited for the mapping generation. This kind of cluster indicates that there are attributes in the data warehouse schema which are very close and may, perhaps, be fused into a unique table.

Running example The real scenario we refer to is an ongoing experiment within the LISEA lab project, funded by Italian Emilia Romagna region. It concerns the ETL process for the creation of a data warehouse in the field of beverage and food logistics software. A new system, called Bollicine Community business Intelligence (BCI), has been proposed to a consortium of companies for: 1) the analyzing and planning the enterprise market starting from its past data; 2) developing a performance benchmarking with respect to general indexes (KPI) obtained by aggregating data of all the members. To reach this goal it is necessary to load all the data about the consortium members in

[3] http://wordnet.princeton.edu/

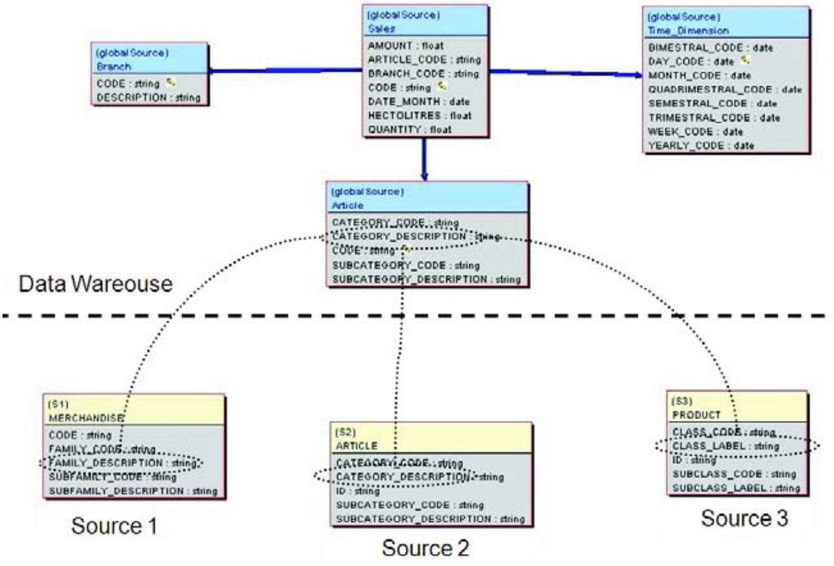


Fig. 2. The mappings for the attribute CATEGORY_DESCRIPTION

the BCI data warehouse. We experimented our tool in this context: preliminary qualitative results show that the enterprises considerably save human resources. In order to explain the approach, we describe the process for inserting into the data warehouse the data of three new companies. The result is that correct mappings has been found, thus improving time and result quality of the extraction process. In particular, let us consider the SALES fact table of the data warehouse, consisting of 3 analysis dimensions (ARTICLE, BRANCH, TIME) (see figure 2):

Let us focus on the attributes FAMILY_DESCRIPTION, CATEGORY_DESCRIPTION, CLASS_LABEL of the new sources. The designer may set the tool in order to take into account only the syntactic similarity measure, thus obtaining two clusters, one made of the FAMILY_DESCRIPTION(S1), CATEGORY_DESCRIPTION(S2), CATEGORY_DESCRIPTION(DW) attributes and the second one with only the CLASS_LABEL attribute.

Since the attributes FAMILY_DESCRIPTION(S1), CATEGORY_DESCRIPTION(S2), CATEGORY_DESCRIPTION(DW) are annotated with the same ‘‘description” concept in WordNet and CLASS_LABEL is annotated with “label” that is a hyponym term of ‘‘description” in WordNet, the knowledge engine generates this set of SYN relationship in the thesaurus.

```
ARTICLE.CATEGORY_DESCRIPTION(DW) SYN MERCHANDISE.FAMILY_DESCRIPTION(S1)
ARTICLE.CATEGORY_DESCRIPTION(DW) SYN ARTICLE.CATEGORY_DESCRIPTION(S2)
ARTICLE.CATEGORY_DESCRIPTION(DW) BT PRODUCT.CLASS_LABEL(S3)
ARTICLE.CATEGORY_DESCRIPTION(S2) SYN MERCHANDISE.FAMILY_DESCRIPTION(S1)
ARTICLE.CATEGORY_DESCRIPTION(S2) BT PRODUCT.CLASS_LABEL(S3)
MERCHANDISE.FAMILY_DESCRIPTION(S1) BT PRODUCT.CLASS_LABEL(S3)
```

These relationships may be exploited with the semantic similarity thus obtaining a unique cluster with all the attributes and consequently a set of mappings between the DW CATEGORY_DESCRIPTION and the corresponding attributes in the new sources.

3.2 A Transformation Function Based on *RELEVANT* Values

Transformation functions are typically implemented for homogenizing attribute values from different sources, thus allowing users to compare values and to perform data analysis processes. Several functions are proposed by commercial ETL tools to transform numeric data-types on the basis of mathematical functions. Very few transformation functions have been developed for string data-types. Such functions are frequently based on concatenations of different attribute values, and on small syntactic changes, such as case modifications and stemming operations.

We aim at providing a new kind of transformation function based on semantic analysis for string values. By means of this function, semantically related values of a chosen attribute in the new data source and the correspondent values into the data warehouse target are grouped, thus providing a semantic reconciliation of the attribute values. The transformation function is based on relevant values and on a temporary mapping table stored in a staging area for preserving the original values, thus allowing precise data analysis. The transformation function works according to the following steps:

- Attribute domains analysis. *RELEVANT* is used for evaluating if the domains of the new source and the data warehouse attribute are compatible, i.e. they describe similar properties for the attributes. The evaluation is based on the idea that for similar domains, overlapped relevant values are computed. The overlapping degree shows the compatibility of the domains.
- If the domains are not compatible, the user may select to transform and load into the target attribute only the synthesized set of values represented by the relevant values. In this case a temporary mapping table is built for preserving the possibility of more accurate data analysis. The advantage of such solution is a reduced cardinality and consequently a more synthetic representation of the object with a reduced loss of semantics.
- If the domains are compatible, it is possible to completely replace the attribute domain with its relevant values. By means of this function, the values of such attribute, typically an analysis dimension, will consist of homogeneous values thus allowing OLAP analysis (i.e. drill down) on them.

Running example. Let us consider the BCI table describing articles sold by companies, `ARTICLE(CODE,DESCR,CATEGORY_DESCRIPTION,SUBCATEGORY_DESCRIPTION)`, where `CATEGORY_DESCRIPTION` is a dimension for BI analysis. Three sources are involved in the data warehouse, and they use different values to describe similar article categories. The following fragment shows some of the attribute values:

```
- Source: #1 Attribute: FAMILY_DESCRIPTION
  Values: {NOT DEFINED, WATER, BEER, WINE, SOFT DRINK, FOOD)
- Source: #2 Attribute: CATEGORY_DESCRIPTION
  Values: {HARD LIQUOR, BOTTLE WINE, NOT DEFINED, JUICE DRINK, MINERAL
  WATER, BOTTLE BEER, SEVERAL ALCOHOLIC BEVERAGE)
- Source: #3 Attribute: CLASS_LABEL
  Values: {NOT DEFINED, MINERAL WATER, BEER, WINE, ALCOHOLIC DRINK,
  FOOD, BOOZE, FOOD CATERING)
```

The *RELEVANT* application to these values defines a set of 8 clusters whose names are loaded into the data warehouse attribute instead of the 22 original values:

```
RV1: NOT DEFINED {NOT DEFINED}
RV2: WATER {MINERAL WATER, WATER}
RV3: FOOD {FOOD, FOOD, FOOD CATERING}
RV4: SEVERAL {SEVERAL}
RV5: BEER {BEER, BOTTLE BEER}
RV6: WINE {WINE, BOTTLE WINE}
RV7: ALCOHOLIC DRINK {ALCOHOLIC DRINK, ALCOHOLIC BEVERAGE, HARD LIQUOR, BOOZE}
RV8: SOFT DRINK {SOFT DRINK, JUICE DRINK}
```

The original values are stored in a temporary table and they may be exploited to update and revise the values of the data warehouse attribute.

4 Conclusion and Future Work

In this paper we proposed to couple and extend our previous research on data integration and data analysis for creating an ETL tool. In particular, we focused our work on the extraction phase, by implementing a technique that semi-automatically defines mappings between a data warehouse schema and a new data source, and on the transformation phase, by proposing a new function based on relevant values, particularly useful for supporting drill down operations. We experimented our approach on a real scenario, thus obtaining qualitative results on the effectiveness of the approach.

Future work will be addressed on identifying a benchmark and a set of measures in order to perform a complete technique evaluation.

References

1. IBM data warehouse manager. Technical Report. IBM, http://www-3.ibm.com/software/data/db2/datawarehouse/
2. Oracle database data warehousing guide 11g release 1 (11.1). Technical Report Oracle, http://www.oracle.com/pls/db111/portal.portal_db?selected=6
3. SQL Server 2005 Integration Services (SSIS). Technical Report. Microsoft, http://technet.microsoft.com/en-us/sqlserver/bb331782.aspx
4. Beneventano, D., Bergamaschi, S., Guerra, F., Vincini, M.: Synthesizing an integrated ontology. IEEE Internet Computing, 42–51 (September-October 2003)
5. Bergamaschi, S., Po, L., Sorrentino, S.: Automatic annotation in data integration systems. In: Meersman, R., Tari, Z., Herrero, P. (eds.) OTM-WS 2007, Part I. LNCS, vol. 4805, pp. 27–28. Springer, Heidelberg (2007)
6. Bergamaschi, S., Sartori, C., Guerra, F., Orsini, M.: Extracting relevant attribute values for improved search. IEEE Internet Computing 11(5), 26–35 (2007)
7. Eckerson, W., White, C.: Evaluating ETL and data integration platforms. Technical Report. The Data Warehousing Institute (2003)
8. Luján-Mora, S., Vassiliadis, P., Trujillo, J.: Data mapping diagrams for data warehouse design with UML. In: Atzeni, P., Chu, W., Lu, H., Zhou, S., Ling, T.-W. (eds.) ER 2004. LNCS, vol. 3288, pp. 191–204. Springer, Heidelberg (2004)
9. Skoutas, D., Simitsis, A.: Ontology-based conceptual design of ETL processes for both structured and semi-structured data. Int. J. Semantic Web Inf. Syst. 3(4), 1–24 (2007)
10. Vassiliadis, P., Simitsis, A., Georgantas, P., Terrovitis, M., Skiadopoulos, S.: A generic and customizable framework for the design of ETL scenarios. Inf. Syst. 30(7), 492–525 (2005)

Knowledge Source Discovery: An Experience Using Ontologies, WordNet and Artificial Neural Networks

M. Rubiolo[1], M.L. Caliusco[2], G. Stegmayer[2], M. Gareli[1], and M. Coronel[1]

[1] CIDISI-UTN-FRSF, Lavaise 610, Santa Fe, Argentina
[2] CONICET, CIDISI-UTN-FRSF, Lavaise 610, Santa Fe, Argentina
mrubiolo@santafe-conicet.gov.ar

Abstract. This paper describes our continuing research on ontology-based knowledge source discovery on the Semantic Web. The research documented here is focused on discovering distributed knowledge sources from a user query using an Artificial Neural Network model. An experience using the Wordnet multilingual database for the translation of the terms extracted from the user query and for their codification is presented here. Preliminary results provide us with the conviction that combining ANN with WordNet has clearly made the system much more efficient.

1 Introduction

The web grows and evolves at a fast speed, imposing scalability problems to web search engines [1]. Moreover, another ingredient has been recently added: data semantics represented by means of ontologies [2]. Ontologies have shown to be suitable for facilitating knowledge sharing and reuse. Thus, the new *Semantic Web* allows searching not only information but also knowledge. The knowledge source discovery task in such an open distributed system presents a new challenge due to the lack of an integrated view of all the available knowledge sources [3].

The web of the future will consist of small highly contextualized ontologies developed with different languages and different granularity levels [4]. The distributed development of domain-specific ontologies introduces another problem: in the Semantic Web many independently developed ontologies co-exist describing the same or very similar fields of knowledge. This can be caused, among other things, by the use of different natural languages (Paper vs. Artículo), different technical sublanguages (Paper vs. Memo), or the use of synonyms (Paper vs. Article). That is why, ontology-matching techniques are needed, that is to say, semantic affinity must be identified between concepts belonging to different ontologies [2].

In this work, we propose an ANN-based ontology-matching model, and the use of WordNet for codifying terms as an appropriate domain data representation within the ANN-based model. The main contribution of this paper is to share

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 66–73, 2009.

with the community the results of an experience in: a) using WordNet multilingual corpus to codify the domain data, which is useful for improving a traditional web search by considering (indirectly) terms synonyms and translation into different languages; and b) using this appropriate codified data to achieve the benefits of the application of an ANN-based ontology-matching model.

The paper is organized as follows. In section 2, the knowledge source discovery task is explained. Section 3 presents the proposed ANN-based ontology-matching model in detail. The results of the model evaluation and comparison against an ontology-matching algorithm called H-Match as well as a discussion of the experiments are shown in Section 4. Finally, section 5 presents the conclusions.

2 Knowledge Source Discovery: A Motivating Scenario

In open distributed systems such as the Semantic Web, several nodes (domains) need resources and information (i.e. data, documents, services) provided by other domains in the net. Such systems can be viewed as a network of several independent nodes having different roles and capacities. In this scenario, a key problem is the dynamic discovery of knowledge sources that, in a given moment, respond well to the requirements of a node request [3].

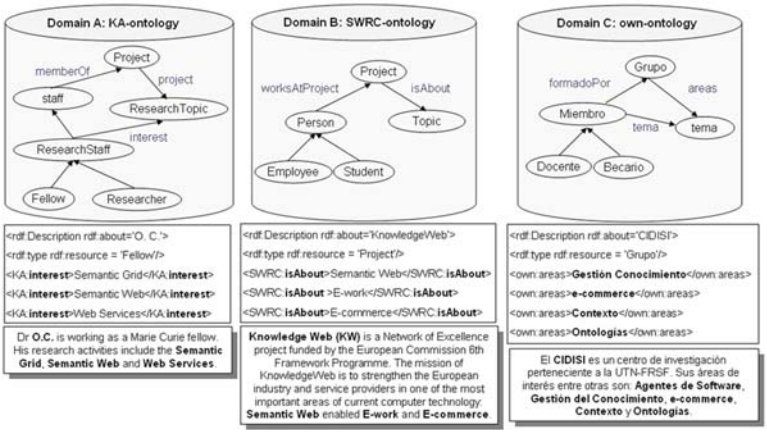
In [5], an architecture for discovering knowledge sources on the Semantic Web was proposed, composed by mobile agents, the Knowledge Source Discovery (KSD) agent and the domains. The mobile agents receive the request from the user and look for an answer visiting the domains according to a list generated by the KSD. The KSD agent has the responsability for knowing which domains can provide knowledge inside a specific area, and it indicates a route to mobile agents that carry a user request. The KSD agent knows the location (url) of the domains that can provide knowledge, but it does not provide the knowledge nor the analysis of what the domain contains (files, pictures, documents, etc.).

The other components of the architecture are the domains. Each domain has its own ontology used to semantically markup the information published in their websites. Suppose there are three domains (A, B, and C) which belong to the Research & Development field of knowledge (figure 1). The domain A uses the KA-ontology[1]. The domain B uses the SWRC ontology[2]. Finally, the domain C uses an own highly-specialized model. As can be seen, each domain may use a different ontology to semantically annotate the provided information even if they belong to the same field of knowledge.

Resource Description Framework (RDF) is used to define an ontology-based semantic markup for the domain website. Each RDF-triplet assigns entities and relations in the text linked to their semantic descriptions in an ontology. For example, in the domain A, the following RDF-triplets: <`O.C., interest, Semantic Grid`>, <`O.C., interest, Semantic Web`> and <`O.C., interest, Web Services`> represent the research interests of O.C. described in the text.

[1] http://protege.cim3.net/file/pub/ontologies/ka/ka.owl

[2] http://ontoware.org/projects/swrc/



```
<rdf:Description rdf:about='O. C.'>
<rdf:type rdf:resource = 'Fellow'/>
<KA:interest>Semantic Grid</KA:interest>
<KA:interest>Semantic Web</KA:interest>
<KA:interest>Web Services</KA:interest>
```

Dr **O.C.** is working as a Marie Curie fellow. His research activities include the **Semantic Grid**, **Semantic Web** and **Web Services**.

```
<rdf:Description rdf:about='KnowledgeWeb'>
<rdf:type rdf:resource = 'Project'/>
<SWRC:isAbout>Semantic Web</SWRC:isAbout>
<SWRC:isAbout >E-work</SWRC:isAbout>
<SWRC:isAbout>E-commerce</SWRC:isAbout>
```

Knowledge Web (KW) is a Network of Excellence project funded by the European Commission 6th Framework Programme. The mission of KnowledgeWeb is to strengthen the European industry and service providers in one of the most important areas of current computer technology: **Semantic Web** enabled **E-work** and **E-commerce**.

```
<rdf:Description rdf:about='CIDISI'>
<rdf:type rdf:resource = 'Grupo'/>
<own:areas>Gestión Conocimiento</own:areas>
<own:areas>e-commerce</own:areas>
<own:areas>Contexto</own:areas>
<own:areas>Ontologías</own:areas>
```

El **CIDISI** es un centro de investigación perteneciente a la UTN-FRSF. Sus áreas de interés entre otras son: **Agentes de Software**, **Gestión del Conocimiento**, **e-commerce**, **Contexto** y **Ontologías**.

Fig. 1. Domains belonging to the R+D field and their semantic annotations

The KSD agent must be capable of dynamically identifying which domains could satisfy a request brought to it by a mobile agent. This dynamic knowledge discovery requires models and techniques which allow finding ontology concepts that have semantic affinity among them, even when they are syntactically different. In order to do this, the KSD agent has to be able to match (probably different) domain ontologies. To face this ontology-matching problem, we propose the use of an ANN model with supervised learning stored in the KSD agent Knowledge Base and trained (and re-trained periodically) off-line.

The KSD agent must also be capable of understanding the natural-language-based query received from the client, which is translated into an RDF-triplet (this process is out of the scope of this work). The resultant RDF-triplet is codified before entering the ANN-based matching model. A WordNet Corpus, which could be composed of different-languages WordNet databases, is used by the KSD agent for this task. WordNet is a lexical database for the English language [6]. It groups English words into sets of synonyms called *synsets*. Every synset contains a group of synonymous words or collocations (sequence of words that together form a specific meaning); different senses of a word are in different synsets. The meaning of the synsets is further clarified by short defining glosses. Most synsets are connected to other synsets via a number of semantic relations that vary according to the type of word, and include synonyms, among others. This research uses WordNet[3] 1.6 since different wordnets for several languages (such as Spanish are structured in the same way.

3 Ontology-Matching: ANN-Based Model and Training

This section presents, through an example, the proposed neural network model for ontology matching and its training strategy.

[3] http://wordnet.princeton.edu/

3.1 The ANN-Based Model

ANNs are information processing systems inspired by the ability of the human brain to learn from observations and to generalize by abstraction. Knowledge is acquired by the network through a learning process, and the connection strengths between neurons, known as synaptic weights, are used to store this knowledge [7].

For neural networks, a matching problem can be viewed as a classification problem. Our ANN-based matcher uses schema-level information and instance-level information (RDF-triplet instances belonging to the RDF annotations of the ontology domain) inside the X ontology to learn a classifier for domain X, and then it uses schema-level information and instance-level information inside the Y ontology to learn a classifier for domain Y. It then classifies instances of Y according to the X classifier, and vice-versa. Hence, we have a method for identifying instances of X $\bigcap$ Y. The same idea is applied to more than two domains.

For building a classifier for Domain X, its RDF-triplets are extracted. Each part of the triplet corresponds to an input unit for the neural model. This way, the proposed model has 3 inputs, each input corresponding to each triplet component. The proposed model is a multilayer perceptron (MLP) neural network model. The outputs of the model are as much neurons as domains. For example, having domain X, Y and Z, the ANN-model has 3 output neurons. The first neuron will be activated each time a RDF-triplet belonging to the domain X is presented to the model. The second neuron will be activated each time a RDF-triplet belonging to the domain Y is presented to the model, and so forth. The activation of a neuron consists in producing a value of (near) 1 when the input RDF-triplet exists in the corresponding ontology domain, and 0 otherwise.

3.2 Training Data

For training the ANN-based ontology matching model, training examples must tell the network that a certain RDF-triplet can be found in a certain domain. This is done through training patterns, which must be numbers.

Once the RDF-triplets are identified for each domain, they have to be codified from string to numbers. To do this, we propose the use of the WordNet database, where a term is associated with a code named *synset offset*. This code is represented by an 8 digit decimal integer. In this way, an appropriate pattern-codification schema can be achieved because all terms can be codified with an invariant-length code.

However, most of the terms represented in WordNet are single words, not collocations. For example, in English WordNet 1.6 the term *Semantic Web* is not a collocation. This is a problem for term codification that is addressed assuming the collocation as two independent words. Then, a triplet term can be represented as a pair of codes, whose values will vary if the term has a) *a single word:* the term code will be formed by the *synset* code associated with the word, including an 8-zero-code in the first position, representing the absence of another word; b) *two words:* the term code will be formed by the composition of a *synset* code associated with each word.

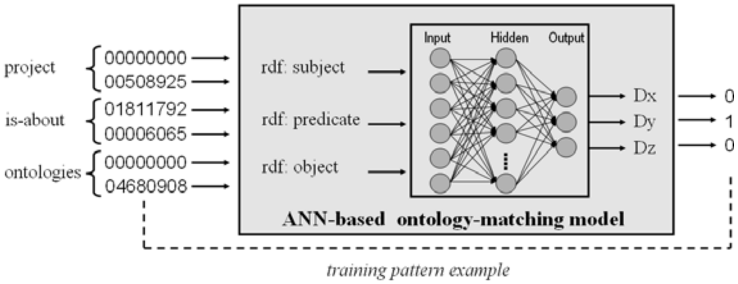


Fig. 2. ANN-based ontology-matching model and its training patterns example

The proposed model uses the standard backpropagation algorithm for supervised learning, which needs {input/output target} pairs named *training patterns* [8]. They are formed by showing to the model, during training, an input pattern of the form: `InputPattern = <rdf:subject; rdf:predicate; rdf:object>`, with its corresponding target value, indicating to which domain is belongs: `OutputPattern = <Dx; Dy; Dz>` (see Figure 2).

The training data are normalized into the activation function domain of the hidden neurons, before entering the model, since this significantly improves training time and model accuracy.

3.3 Training Example

A simple example of one training pattern is presented in figure 2. Considering the ontologies of figure 1, a training pattern indicating that the triplet <*project; is-about; ontologies*> can be found on the Domain B ontology but not on A or C is: `InputPattern=<project;is-about;ontologies>` and `TargetPattern=<0;1;0>`.

This means that, given the fact that there are projects in the domain B whose research interest is about ontologies, its corresponding triplet would be `<project; is-about; ontologies>` and its corresponding output target would be `<0; 1; 0>`: only the second vector value (that represents Domain B) is equal to 1, indicating that this triplet can be found on domain B ontology.

The Figure 2 shows also the triplet codification. The code related to *is-about* is formed by *is* code `01811792` and *about* code `00006065`: <`01811792; 00006065`>. For *project* the code is formed as a combination of zero and `00508925`: <`00000000; 00508925`>. In summary, this training pattern would be:
`InputPattern = <<00000000; 00508925>; <02579744; 00006065>; <00000000; 04680908>>` and `TargetPattern = <0;1;0>`.

An interesting fact related to the use of different language ontologies arises as a consequence of using the WordNet database for triplet term codification. Because all of the words and their translations are codified with the same code in the WordNet database, the process of identifying the right domain for a particular triplet can be significantly improved. There is some sort of automatic triplet expansion and translation as a consequence of using the WordNet codification scheme for ANN model training. That is to say, a term in a triplet has the same code in English WordNet as well as in Spanish WordNet. For example, the

English term *project* and its Spanish translation *proyecto* has the same code: `00508925`. Using this unique code, it is possible to consult all domains, without taking each domain language into account. Similar conclusions can be drawn in the case of synonyms.

4 Evaluating the Proposed Strategy against H-Match

Results of an experience using the ontologies shown in Section 2, the WordNet use for codifying triplet terms from the original queries, and the ANN-based ontology-matching model application are reported in this section.

One very important aspect of evaluation is the data set used for the testing. Datasets for matching ontologies are not easy to find. One problem is that they require public and well-designed ontologies with meaningful overlap. The data sets made for OAEI [4] (Ontology Alignment Evaluation Initiative) campaign can be considered as correct by construction but they are not realistic nor very hard. In addition, all data sets defined for evaluating matching algorithm are composed by a pair of ontologies. In contrast, to evaluate the proposed method more than two ontologies are required[9].

The MLP model parameters are set according to typical values, randomly initialized. The number of input neurons for the MLP model is set to 6, considering a double-code for each triplet term. The hidden layer neuron number is set empirically, according to the training data and the desired accuracy for the matching. At the output, there is a specialized output neuron in the model for each domain. The allowed values for each output neuron are 1 or 0, meaning that the neuron recognizes or not a concept belonging to the domain it represents.

The ANN-based ontology-matching model is trained with each domain ontology RDF-annotations and their corresponding instances. Since we need a populated ontology, we have semantic annotated three different web pages obtaining 134 patterns for the ANN model training.

It is difficult to make a comparison of our proposal against others matching algorithm due to the fact that most matching algorithms work on structured ontologies. That is to say, as we need a populated ontologies to train our model, we cannot apply any of the matching algorithms to the problem because they are focused on structured ontologies evaluation.

4.1 Comparison Results with H-Match Algorithm

The proposed ANN-based ontology-matching model has been compared with the H-Match [10], an algorithm for matching populated ontologies by evaluating the semantic affinity between two concepts considering both their linguistic and contextual affinity.

In order to use this algorithm, a probe query (one word) is sent to each domain, which applies the algorithm to determine whether it has concepts matching it or not. The six examples are later evaluated domain by domain setting

[4] http://oaei.ontologymatching.org/

Table 1. Ontology-matching results comparison

Query	*Domain*	*ANN − model*	*H − Match*
1) <fellow,interest,semanticWeb>	A	A	A
2) <miembro,tema,gobierno>	C	C	A,C
3) <project,is-about,ontologies>	A,B	A,B	A,B
4) <researcher,topic,web>	B	C	A,B,C
5) <-,-,semanticGrid>	A	A	A,B,C

the algorithm parameters as: matching model = *intensive*, mapping = *one-to-one*, adopts inheritance = *false*, empty context strategy = *pessimistic*, matching strategy = *standard (asymmetric)* and weight linguistic affinity = *1.0*.

The H-Match algorithm provides a semantic affinity value ($S_{i,D}$) for each triplet-term i compared with each domain ontology D. These values are combined to obtain an average matching measurement (Av) for each complete triplet $t_{i,j,k}$ against a domain ontology, according to $Av_{(t_{i,j,k},D)} = \frac{S_{i,D}+S_{j,D}+S_{k,D}}{3}$. To determine if the triplet can be indicated as belonging or related to the analyzed domain D, the semantic affinity measurement $Av_{(t_{i,j,k},D)}$, as well as two of the semantic affinity values, have to be higher than an empirically set threshold of 0.7. If both conditions are satisfied, the triplet $t_{i,j,k}$ is considered to be "matched" to the domain ontology D.

The results of the analysis of the ANN-based model against H-Match algorithm are reported in Table 1. The first column indicates the triplet query considered in the test and the second column indicates which domain it should be associated with. The third column reports the results obtained from the use of the proposed ANN-model for the ontology-matching task, while the fourth column reports the results from the use of the H-Match algorithm.

From the results shown in Table 1, it can be stated that the proposed model can be quite accurate for indicating potential domains that can answer a query, compared to a traditional matching algorithm. Note that the query triplet 3) <*project,is-about,ontologies*> has a translation in both domain A and domain B ontologies, and in fact the ANN model indicates that the domain ontologies of A and B contain some ontology labels or instances that are similar to the presented request.

Another interesting test queries are 2) <*miembro, tema, gobierno*> and 4) <*researcher, topic, web*>. As can be noted, the two first triplets components are translations of the same words and the ANN-based model provides the same answer for both cases, showing the advantage of using a codification scheme for words which is independent of the language. However, here the neural model shows a flaw: the domain C is indicated as the final result because it is the last domain examples the ANN model has seen during the training process. It is an indication that some procedure must be used during training for assuring model independency of trainingg patterns order, such as bootstrapping, cross-validation or leave-one-out algorithms. For all the remaining tests, the ANN-based ontology-matching model has provided satisfactory results.

5 Conclusions

The ontology-based knowledge source discovery on the Semantic Web, focused on discovering distributed knowledge sources from a user query using ANN models and WordNet multilingual database, was experienced in this paper. Using the Wordnet multilingual database for codifying the triplet terms, extracted from the user query, some sort of automatic triplet expansion and translation arose, which improved the traditional search task. This codification also allowed appropriately representing the ANN-based ontology-matching model input data. This paper has shown the benefits of including an ANN-based ontology-matching model inside a KSD agent, whose capabilities for discovering distributed knowledge sources have been improved. In addition, the combination of ANN with WordNet has clearly made the system much more efficient.

References

1. Baeza-Yates, R.: Web mining. Proceedings of LA-WEB Congress 1(2), 19–22 (2005)
2. Davies, J., Studer, R., Warren, P.: Semantic Web Technologies: trends and research in ontology-based systems (2007)
3. Castano, S., Ferrara, A., Montanelli, S.: Dynamic Knowledge Discovery in Open. Distributed and Multi-Ontology Systems: Techniques and Applications (2006)
4. Hendler, J.: Agents and the semantic web. IEEE Intelligent Systems 16(2), 30–37 (2001)
5. Stegmayer, G., Caliusco, M., Chiotti, O., Galli, M.: ANN-agent for Distributed Knowledge Source Discovery. On the Move to Meaningful Internet Systems (2007)
6. Fellbaum, C.: WordNet: An Electronic Lexical Database (Language, Speech, and Communication). The MIT Press, Cambridge (1998)
7. Haykin, S.: Neural Networks: A Comprehensive Foundation (1999)
8. Wray, J., Green, G.: Neural networks, approximation theory and precision computation. Neural networks 8(1), 31–37 (1995)
9. Euzenat, J., Shvaiko, P.: Ontology Matching. Springer, Heidelberg (2007)
10. Castano, S., Ferrara, A., Montanelli, S., Racca, G.: Semantic information interoperability in open networked systems. In: Bouzeghoub, M., Goble, C.A., Kashyap, V., Spaccapietra, S. (eds.) ICSNW 2004. LNCS, vol. 3226, pp. 215–230. Springer, Heidelberg (2004)

Path Planning Knowledge Modeling for a Generic Autonomous Robot: A Case Study

Rafael Guirado, Clara Marcela Miranda, and José Fernando Bienvenido

Departamento de Lenguajes y Computación, University of Almeria,
Ctra. Sacramento, S/N, E-04120, Almería, Spain
{rguirado,cmiranda,fbienven}@ual.es

Abstract. This paper presents the initial steps followed in order to build an ontology about robot navigation (including specifically alternative navigation algorithms). Tackling the problem from general to specific, we start analyzing the desired behavior for generic mobile robots, in order to get common tasks and methods. Then, we fix our attention into the agricultural spraying robot developed in the University of Almeria by a multidisciplinary team. Because the field of robot navigation is consolidated, there are many algorithms (methods) to perform same activities (tasks). Our goal is to build an ontology including all this alternative methods, applying the dynamic selection of methods to make decisions in real-time depending on the environment conditions. Here we show the task-method diagrams with the parameterized description of some of alternative methods, using Fitorobot as testing case.

Keywords: Knowledge management, Ontology, Dynamic selection of methods, Path planning, Spraying robots.

1 Introduction

For a long time, research centers and universities have been looking how to facilitate human work (releasing humans of the most dangerous tasks), applying robotic techniques. In this sense, and given that most of the tasks require human displacements, it is required to design robots that incorporate navigation algorithms. One of the main income sources of the Province of Almeria is agriculture. In this context, a multidisciplinary research group of the University of Almeria is working in the design of a mobile robot (called *Fitorobot*) that permits movement between lines of crop, and the performance of several greenhouse tasks such as spraying, pruning, and crop transport (see Fig. 1).

Initially, in order to control the robot navigation we considered two different approaches: a map-based deliberative technique and a pseudo-reactive technique. These approaches could be integrated as a two levels decision tree. At the first level it is fixed the task to be performed in order to reach the final objective (when it is required this decision). In the second level, it is selected the best method in order to achieve the objectives of the task (different methods suppose alternative ways to reach the task goal). The general idea is to assemble a navigation model of the mobile robot as

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 74–81, 2009.

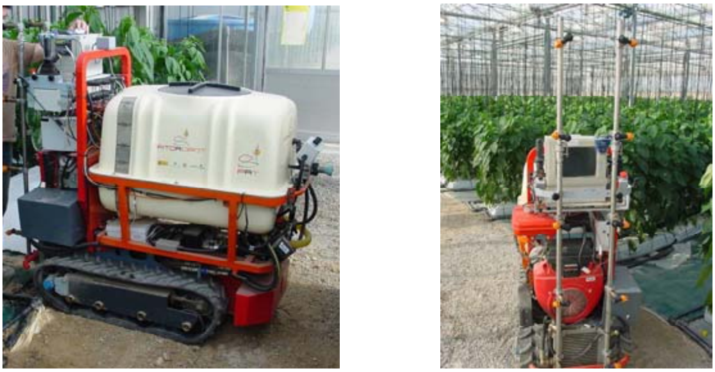


Fig. 1. Lateral view and back view of the mobile robot *Fitorobot* into a greenhouse

general as possible, deciding between the different navigation alternatives on the fly. In order to do this, we must build a general ontology of navigation algorithms (and their components), that including all the tasks and methods, would constitute a battery of parameterized navigation elements as general and complete as possible.

2 System Description

Our testing robot, *Fitorobot,* has a differential-drive mechanism of locomotion. The system is composed of two rubber-tracks, which provide a larger contact surface with the soft ground of the greenhouses, making it more robust and stable. This robot has a mass of 756 kg (with the spray tank full), and it has appropriate dimensions for the typical corridors of greenhouses in south-eastern Spain. It is driven by a 20-hp gasoline engine. It also has a low-cost sensor system, including low-distance ultrasonic sensors, middle-distance ultrasonic sensors, magnetic compasses, incremental encoders, radars, and security sensors (see Fig. 2a). Furthermore, for spraying control, a pressure sensor has been installed.

About the navigation techniques, it starts evaluating if it is accessible a map of the greenhouse; in this case it applies a deliberative method. On the other hand, when there is no map, a pseudo-reactive algorithm is used. Moreover, a sensorial map is built along the path, to be employed by the deliberative method in later runs. The two previous approaches utilize a security layer to avoid collisions; this layer uses on/off sensors. Finally, it has a low-level (servo) control layer composed of two PID controllers that regulate the speed of the tracks. Fig. 2b shows the navigation strategy schema for the mobile robot proposed in [1] by R. González et al.

Greenhouses are structured environments where the distribution of plants is at least partially known. The main obstacle to the movement of mobile robotics in greenhouses is related to the fact that navigation algorithms should take into account unexpected events (humans working in the greenhouse). Furthermore, appropriate filters for the sensor readings, and robust navigation strategies should be examined.

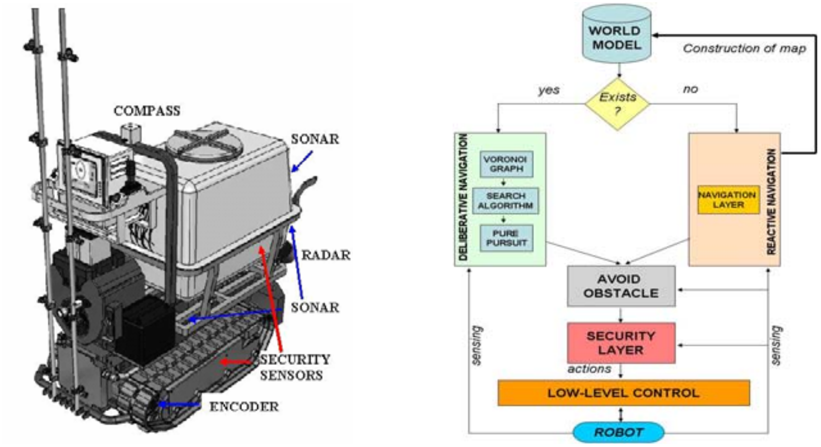


Fig. 2. *Fitorobot*: a) Sensorial system and b) Navigation strategy implementation (after the integration of the different methods, including lower level security)

3 Methodology

The knowledge model, about the navigation of mobile robots described in this works, was assembled using some elements of the CommonKADS methodology and the dynamic selection of methods (DSM). Now, we are going to introduce those techniques and a short resume of the navigation algorithms included in the system. This knowledge model, despite it has been developed for Fitorobot, is general.

3.1 Path-Planning Algorithms (Methods)

From the highest point of view, planning the movements of a mobile robot can be done using two main approaches: global (deliberative) techniques and local (reactive) techniques. First group requires the knowledge of the robot environment, second

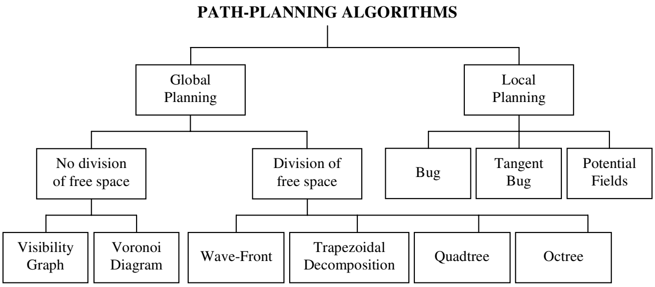


Fig. 3. Common algorithms for path planning (as described in [2])

group do not. In our work case, as the robot is moving into a given environment, a greenhouse, it would be possible to apply the global approach; but, some alterations (as a box in the path) force the application of a combination of map based algorithms with reactive ones.

Fig. 3 shows most usual movement planning algorithms, classified following the previously given dichotomy: global and local planning algorithms. [2] offers a short description of these algorithms with a tool that let us to observe their behavior in different situations in a fully interactive way. This way it has been possible to evaluate the suitability of the different methods for the specific situations to be used in the process of dynamic selection of planning methods.

3.2 CommonKADS Methodology

The CommonKADS methodology was consolidated as a knowledge engineering technique to develop knowledge based systems (KBS) at the beginning of the 90's [3]. It includes a kernel set of models, which are summarized in Fig. 4a. In our case, we have worked in the knowledge model and specifically in the assembling of tasks, methods, inferences and domain knowledge elements, as presented in Fig. 4b.

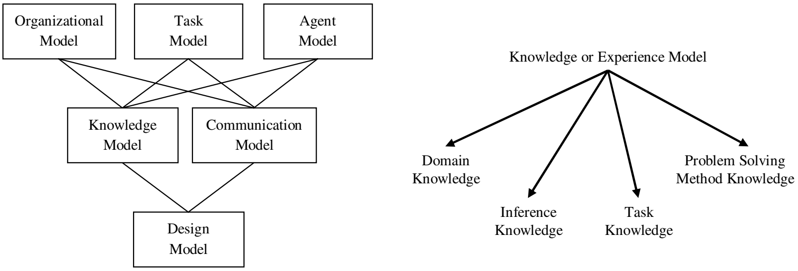


Fig. 4. CommonKADS models: a) Set of models, b) Knowledge model components

The most representative tools to model the problem solving mechanisms are the Task-Method Diagrams (TMD), that present the relation between one task to be performed and the methods that are suitable to achieve this task, followed by the decomposition of these methods in subtasks, transfer functions and inferences (final implemented algorithms), as shown in a general way in Fig. 5. The main problem to be solved is represented by the highest level task [4].

3.3 Dynamic Selection of Methods

A given task can be achieved by more than one method, and these can be applied only in specific environmental conditions. This way, it would be required that the robot selects one of the suitable methods on the fly (using data received from its sensors).

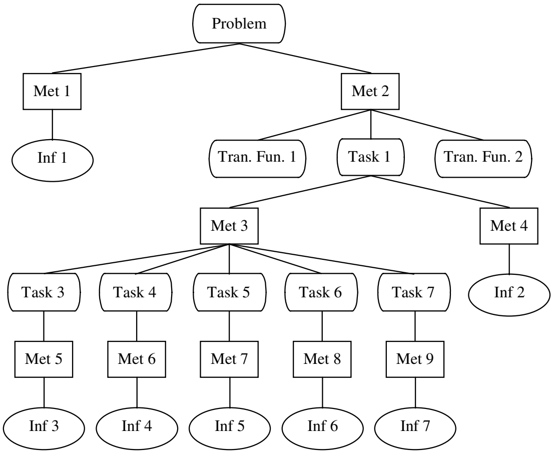


Fig. 5. Simple TMD

Table 1. Suitability criteria of 4 alternative methods (for path-planning)

Algorithm	Minimum distance	Computing time	Obstacle proximity
Wave-Front	1	1	3
Voronoi Diagram	2	2	1
Cell Decomposition (Quadtree)	2	2	4
Irregular Cell Decomposition	1	3	4

We proposed to assemble a general decision module that, taking account of the suitability criteria defined for each alternative method and actual data, would activate the most adequate method. These suitability criteria are assigned weights whose values can be update both manually and automatically [5]. Table 1 shows the main suitability criteria for the methods that compute the most adequate path for the spraying task in a partially known greenhouse. The cost function considers the three criteria, using a higher weight for the third one (it is the most relevant); when there are near obstacles are preferable last two methods, if we have enough computing time. This technique was used previously in configurational design of greenhouses and aeronautic conformation pieces as described in [5].

4 Modeling the Navigation System

Modeling the knowledge, we try to explicit all the knowledge implicit in the texts written by field experts (in our case, the navigation of mobile robots). Assembling these models required of an intensive bibliographical evaluation.

We propose a generic knowledge model about the navigation of robots, based in the CommonKADS methodology (mainly some modeling tools) and the dynamic selection of methods, to be used by the navigation subsystems of mobile robots. From a general point of view, within this area, we can find a huge casuistic, given different environment characteristics (as the different obstacles) and the kinematic, dynamic and sensorial characteristics of the robot. As an extreme example, we could find different types of obstacles less o more "dangerous" for the robot (e.g. a less stable objects could easily fall on the robot when contacting it, causing different damages). For this reason, the knowledge model must include a very detailed specification of each one of the elements involved in the navigation tasks, in order to offer the robot the most appropriate alternative for each particular situation.

This model of knowledge starts from a global problem, that it is navigating in its specific environment in order to achieve a specific objective (this is therefore the main task of our model). Achieving this general objective, under the specific conditions of the system (in this level the availability of a map of the environment, that means that it is known or not), can be done activating two initial alternatives: global or local methods. First group supposes calculating an efficient path using a specific criterion (minimum distance, maximum distance to obstacles,...), rising new tasks (second order tasks). The second group of methods may require the application of an exploration task (recognition of the environment to generate a map) or the application of a task that starts the navigation of the robot detecting and avoiding obstacles on the road. If it is explored the environment, generating a map, the global method previously described would be applicable. As the description of this model is huge, Fig. 6 shows the higher level elements of the associated TMD.

The "Traverse-along-a-Path-Segment" task repeats iteratively the two subtasks that compose it, the "Planning-of-Path-Segment" and "Tracing-Calculated-Path-Segment" tasks, strictly in this order until reaching the goal position or finding that it is unattainable. This iterative execution of a task is represented by an * on the top right of the task symbol.

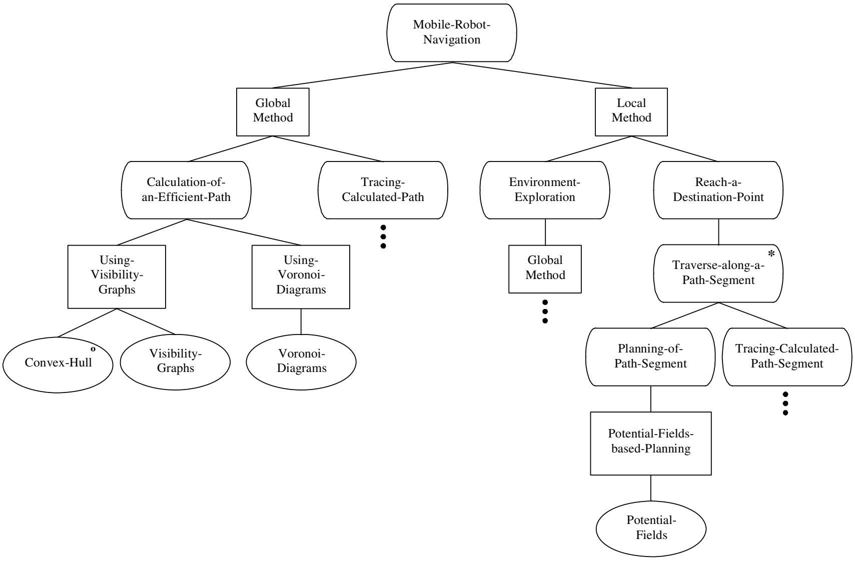


Fig. 6. Partial representation of a TMD for mobile robot navigation

The Convex-Hull inference evaluates the convex boundary for an obstacle or set of obstacles. It is used when the system detects an obstacle or set of obstacles near the robot with any concave vertex in its boundary. As shown in the CML description of this inference, this algorithm is not executed when the robot is located or must access an area inside a concave region of the obstacle (in order to avoid the inclusion of the trajectory inside the virtual object.

Dynamic selection of methods is applied, for example, in the task of calculation of the efficient path. In this case, we show only two alternatives (from the wide set of alternatives found) for the achievement of the task goal; these alternatives have been assigned a set of suitability criteria and associated weights. These suitability criteria and weights let the system to decide which method must be activated.

As proposed in CommonKADS, the different elements (tasks, methods and inferences) of the previously defined TMD are modelled using CML schemas. These schemas formalize all the knowledge associated to each one of these elements. Next, we show some relevant parts of an inference used in the diagram. This is used to model one process (algorithm) that let the robot to reach some partial goal.

A simplified description of the Convex-Hull inference would be:

```
INFERENCE Convex-Hull;
 OPERATION-TYPE: calculate;
 ROLES:
  INPUT:
   visibility-graph:
    "formal description of a graph, that represents all the
    vertices of the obstacles and the initial and final
    configurations; these vertices are matched according to
    their visibility, ie whether the line that connects them
    do not intersect with any other obstacle";
  OUTPUT:
   convex-visibility-graph:
    "formal description of the visibility graph, representing
    the obstacles as convex polygons; if the initial or final
    configuration is included into a concave area of an
    obstacle, this inference is not applicable";
 SPECIFICATION:
    "1. Look for the vertices with higher and lower values of
    x and y (4 vertices).
    2. Order all the vertices by increasing value of x.
    3. Select the vertex with the lower value of x ($x_{min}$). Put
    it in a stack and look for the extremes between vertices
    $x_{min}$ and $y_{min}$.
       a. Compute the slope of all the lines that start from
       the vertex which is at the top of the stack and go to
       all the vertices that are at the left of $y_{min}$ and right
       and below of that located on top of stack.
       b. Select the vertex associated to the line with the
       more negative slope, and put it in the stack.
    4. Repeat step 3 with the new vertex on the stack until
    the vertex corresponding to $y_{min}$ would be put on the stack.
    5. Repeat steps 3 and 4 drawing the lines between the $y_{min}$
    and $x_{max}$ vertices, looking for the vertices on the right of
    $y_{min}$ and on the left and down of $x_{max}$ with minimum positive
    slope.
```

```
      6. Repeat steps 3 and 4 drawing the lines between the $x_{max}$
      and $y_{max}$ vertices, looking for the vertices over $x_{max}$ and
      right and down of $y_{max}$ one, with maximum negative slope.
      7. Repeat steps 3 and 4 drawing the lines between the $y_{max}$
      and $x_{min}$ vertices, looking for vertices down of $y_{max}$ and on
      the right and over the $x_{min}$ one, with minimum positive
      slope"
END INFERENCE Convex-Hull;
```

This way, the CML schemas of all the elements (task, methods, inferences and transfer functions), that configure the TMD for robot navigation, are assembled.

5 Conclusions and Future Works

Main objective of this work was to present a dynamic mechanism to order the different/alternative algorithms for robot navigation. Main advantages were to facilitate further addition of new algorithm that could be developed in the future, and the capacity of deciding on the fly the most adequate to be used in specific conditions (in a general way). We present a general planning system that would decide the most adequate algorithm to be used, selecting one method from our repository of well parameterized methods.

In order to evaluate the proposed mechanisms of dynamic selection of methods in robotics, actually we are preparing the use of these techniques in the field of social (or sociable) robots, where there are much more application opportunities with wider and more complex alternatives.

References

1. González, R., Rodríguez, F., Sánchez-Hermosilla, J., Donaire, J.G.: Navigation Techniques for Mobile Robots in Greenhouses. Applied Engineering in Agriculture 25(2), 153–165 (2009)
2. Guzmán, J.L., López, O., Berenguel, M., Rodríguez, F., Dormido, S., Piguet, Y.: MRIT: Mobile Robotics Interactive Tool. In: Internet Based Control Education IBCE 2004, Grenoble (2004)
3. Schreiber, G., Akkermans, H., Anjewierden, A., de Hoog, R., Shadbolt, N., Van de Velde, W., Wielinga, B.: Knowledge Engineering and Management. The CommonKADS Methodology. MIT Press, Cambridge (1999)
4. Bienvenido, J.F.: Selección dinámica de métodos y reutilización de elementos de conocimiento: Una extensión a CommonKADS aplicada a problemas de diseño. PhD, University of Murcia (1999)
5. Bienvenido, J.F., Flores-Parra, I., Guirado, R., Marín, R.: Knowledge Based Modeling of the Design Process as a Base of Design Tools. In: Moreno-Díaz Jr., R., Buchberger, B., Freire, J.-L. (eds.) EUROCAST 2001. LNCS, vol. 2178, pp. 209–222. Springer, Heidelberg (2001)

System Models for Goal-Driven Self-management in Autonomic Databases

Marc Holze and Norbert Ritter

University of Hamburg, Department of Informatics,
Vogt-Kölln-Str. 30, 22527 Hamburg, Germany
{holze,ritter}@informatik.uni-hamburg.de
http://www.informatik.uni-hamburg.de/

Abstract. Self-managing databases intend to reduce the total cost of ownership for a DBS by automatically adapting the DBS configuration to evolving workloads and environments. However, existing techniques strictly focus on automating one particular administration task, and therefore cause problems like overreaction and interference. To prevent these problems, the self-management logic requires knowledge about the system-wide effects of reconfiguration actions. In this paper we therefore describe an approach for creating a DBS system model, which serves as a knowledge base for DBS self-management solutions. We analyse which information is required in the system model to support the prediction of the overall DBS behaviour under different configurations, workloads, and DBS states. As creating a complete quantitative description of existing DBMS in a system model is a difficult task, we propose a modelling approach which supports the evolutionary refinement of models. We also show how the system model can be used to predict whether or not business goal definitions like the response time will be met.

1 Introduction

For several years, researchers and DBMS vendors have developed self-management functions for particular DBMS components or administration tasks. Thus, commercial DBMS today typically provide an index advisor, autonomic memory management, and automated maintenance functions. These self-management approaches usually follow the feedback control loop (FCL) pattern, where a *controller* continuously monitors a resource via *sensors*, analyses the observed information, and immediately performs necessary reconfigurations via *effectors*.

Already at an early stage of autonomic DBS technology development it has been noticed [1] that adding FCLs to a DBS entails severe problems. These problems are primarily caused by the fact that the FCLs lack knowledge about the components they manage: First, they cannot predict the performance effect of planned reconfigurations on the managed resource. Autonomic memory management, for example, "guesses" the required memory adaptations in small steps of fixed size. This trial-and-error bears the risks of *overreaction*, because DBS reconfigurations may require a long time before the reconfigurations take effect.

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 82–90, 2009.

Second, the FCLs lack knowledge about the system-wide effects of their reconfigurations. For example, an autonomic connection management may increase the number of DB agents to improve the overall system throughput. However, this decision may also affect the performance of other components, as it reduces the memory available for the system buffer. The lack of knowledge about side-effects causes *interference* between the decisions of the FCLs. Third, the relationship between the performance of the managed component and system-wide goals (e.g. response time, availability) is unknown to the FCLs (*goal-independency*). Knowledge about the relationship to system goals would allow the FCLs to restrict expensive reconfiguration analyses to situations when they are actually required.

In this paper we present an approach towards system-wide, goal-driven DBS self-management. In particular, we identify the knowledge that is required to prevent the problems of overreaction, interference and goal-independency. We propose the usage of a system model, which serves as a centralized knowledge base for the self-management logic. By choosing the graphical modelling language SysML for the system model, our approach is designed to be used for existing DBMS. We show how knowledge about both the structure and the behaviour of the DBMS can be expressed in this model, and how the system model can be extended by objective functions that can be evaluated against goal definitions.

The paper is structured as follows: Section 2 discusses the necessary system model contents and the relevant goal types. In Section 3 we describe how a system model can be created for existing DBMS, using IBM DB2 as an example. The objective functions for the goal-driven optimization is subject of Section 4. We discuss related work in Section 5 before we conclude with an outlook in Section 6.

2 Goal-Driven Self-Management

In order to reduce the operation and maintenance costs, DBS should automatically consider the users' performance requirements (goals) during query processing, e.g. by treating queries with high priority preferentially. However, as no existing DBMS provides this functionality, our objective is to create an external self-management logic, which can be applied to existing DBMSs. Hence, meeting the goals must be accomplished by using the existing configuration options of a DBMS only. For this purpose, it is essential for the self-management logic to maintain a system-wide view on the DBS. As shown in Fig. 1, the *workload* and the *state* of the DBS must be continuously monitored by a *self-management logic*. The workload provides information on how a DBS is used in its particular environment (e.g. CPU usage, SQL trace). The state information refers to internal characteristics like the average response time or the accuracy of optimizer statistics. The self-management logic must compare the current state of the DBS to the *goals* defined by a DBA, and start a reconfigurations analysis when there is a risk of missing the goals. For the decision on which reconfigurations will meet the goals under the current workload, the self-management logic needs detailed knowledge about the DBS. We refer to this knowledge as the *system model*.

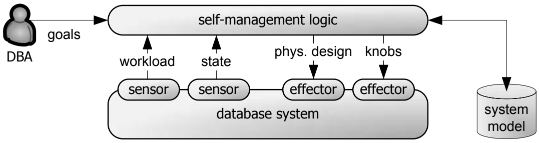


Fig. 1. DBS Self-Management Overview

From the knowledge stored in the system model, it must be possible to decide how the computing resources must be shared amongst the DBMS components. In addition, the self-management logic must be able to assess configuration alternatives (access paths, number of bufferpools, tablespace design, ...) in a specific environment. So the system model must comprise two main parts: a *Hardware Model* describing the available computing resources, and a *DBMS Model* describing the DBMS components and their behaviour. It is the task of self-management logic to evaluate the knowledge stored in the system model in order to decide which reconfigurations are necessary to meet the goals. Hence, the system model must also comprise *goal functions*. These functions must quantitatively describe the values expected for the goals response time, throughput, CPU/disk usage, availability, and operation costs, depending on the DBS configuration.

3 System Model

In the following, we describe our approach for storing the knowledge for DBS self-management in a system model. Section 3.1 first identifies the necessary model contents, before Section 3.2 describes the selected modelling technique.

3.1 Model Contents

As described in Section 2, the system model stores knowledge about the structure of the DBMS and the available hardware. The DBMS model is hierarchically structured (see Fig. 2), where every level refines the component structure. For every component the sensors, effectors and constraints must be defined:

As the *sensor* data must be read and interpreted by the self-management logic, the DBMS model must describe how the sensor value can be retrieved (*availability*, e.g. from system catalogue). Furthermore, the meaning of the sensor information (*semantics*, e.g. counter, or high water mark), and its *type* (workload or state) have to be defined. For *effectors* it is necessary to describe their *type* (e.g. configuration parameter, physical design, maintenance function) and whether they can be manipulated online (*changeability*). The effector values of DBMS components in many cases may not be changed arbitrarily, but are subject to *constraints*. For example, there usually are rules for the allowed values (*domains*). In addition, configuration parameters of one component may depend on configuration values of another component (*dependencies*). The most important constraints for the automatic deduction of reconfiguration actions is the description of the expected *behaviour* of the component. It must be described in terms

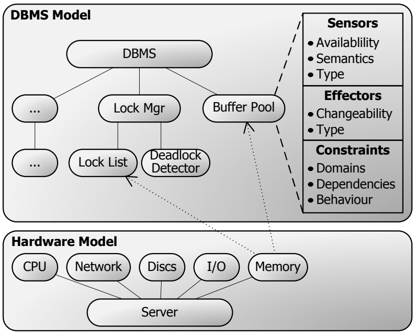


Fig. 2. DBMS- and Hardware-Model

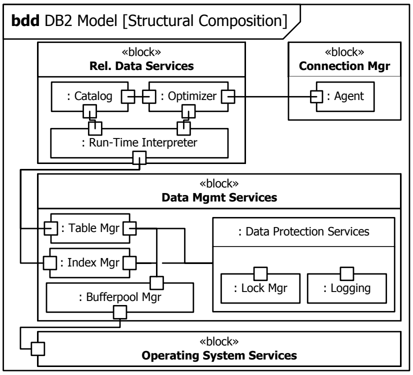


Fig. 3. DB2 Structure Example

of a mathematical model of the component, which quantifies the performance of the component depending on its sensor and effector values. Only then it is possible for the self-management logic to predict the effects of reconfigurations.

It is important to note that the description of the logical and physical design of a DBS is *not* part of the DBMS model. This information is instead available from the system catalogue as sensor information, and it can be adapted via effectors in order to influence the performance of the DBMS components.

The hardware information in the system catalogue usually does not provide information on the performance characteristics and costs of using a particular piece of hardware. However, this information will be required by the self-management logic in order to minimize computation cost, while meeting the performance goals. Hence, a separate hardware model must be maintained. This model will also allow the explicit representation of resource competitions between DBMS components (illustrated as dotted arrows in Fig. 2).

3.2 DBMS System Modelling

Due to the complexity of today's commercial DBMSs, a complete and exact mathematical model of the system behaviour is not feasible. But as noted by Weikum et al. in [1], even an approximate, coarse-grained model would be a step ahead. Hence, our approach towards a system model is to create a model outside of the self-management logic. In contrast to implementing the knowledge about the system structure and behaviour directly in the logic, placing it in a separate model has the advantage that it can be easily extended, refined and adapted. So a very coarse-grained model can be created in a first step, containing the most important components, a subset of their actual sensors and effectors, and an approximated quantification of their behaviour. After this model has proven to sufficiently well predict the system behaviour, it should be refined incrementally.

For the definition of a system model we have chosen the SysML modelling language [2], because it allows the representation of all the required model contents identified in Section 3.1. SysML has been designed to support the specification,

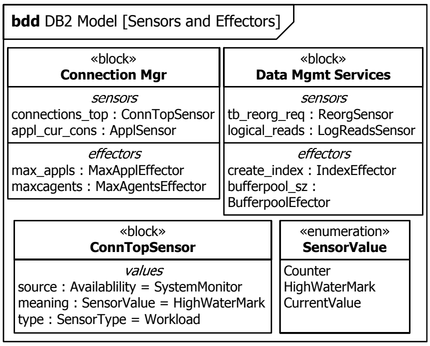

bdd DB2 Model [Sensors and Effectors]
«block»
Connection Mgr
sensors
connections_top : ConnTopSensor
appl_cur_cons : ApplSensor
effectors
max_appls : MaxApplEffector
maxcagents : MaxAgentsEffector
«block»
Data Mgmt Services
sensors
tb_reorg_req : ReorgSensor
logical_reads : LogReadsSensor
effectors
create_index : IndexEffector
bufferpool_sz :
BufferpoolEfector
«block»
ConnTopSensor
values
source : Availablility = SystemMonitor
meaning : SensorValue = HighWaterMark
type : SensorType = Workload
«enumeration»
SensorValue
Counter
HighWaterMark
CurrentValue

Fig. 4. Sensors/Effectors Example

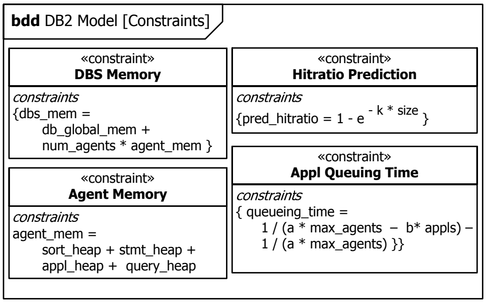


Fig. 5. DB2 Constraints Example

analysis, design, and validation of a broad range of systems and systems-of-systems, including hardware, software, information, and processes aspects [3]. For the definition of the structure of the system as well as the constraints on effector values we use *block definition diagrams* and *internal block diagrams*. For the quantification of the system behaviour we employ *parametric diagrams*.

We illustrate the usage of SysML for the definition of DBS system models by describing an exemplary model for IBM DB2. A coarse-grained structural model of DB2 is given in terms of a SysML block definition diagram in Fig. 3. It shows the hierarchical structure of the main DB2 components: The *Connection Manager* handles the connections to client applications by assigning *Agents* to them, which execute the SQL statements in the DBS for the client. The agent passes the statements to the *Relational Data Services*, where execution plans are created for them using the *Optimizer* and *System Catalog*. These plans are executed in the *Run-Time Interpreter*, where they use the *Table Manager* and *Index Manager* components of *Data Management Services* to retrieve the data. The Table Manager and Index Manager request the necessary pages from the *Bufferpool Manager*, which accesses the files on disk via the *Operating System Services*. Isolation and Durability are ensured by the *Data Protection Services*.

The block diagrams in Fig. 3 are well suited to depict the structural composition of a DBMS. For the definition of the components' sensors and effectors we use an additional block definition diagram. Examples for sensor and effector definitions of DBMS components are given in Fig. 4. The Connection Manager, for example, has a sensor *connections_top* for the highest observed number of client connections. Every sensor and effector has to be represented by a separate block in the model, as specific characteristics must be stored (e.g. *ConnTopSensor*).

In addition to the structural information about DBMS components, the DBMS model has to describe the effectors' domain constraints, dependencies, and the component behaviours. For these purposes the SysML ConstraintBlock element is used. The left part of Fig. 5 shows an example for two constraints *DBS Memory* and *Agent Memory*, which define the dependencies between the sizes of different memory areas. It shows that the memory available in the DBS global memory (e.g. bufferpools) depends on the number of agents. The right part of

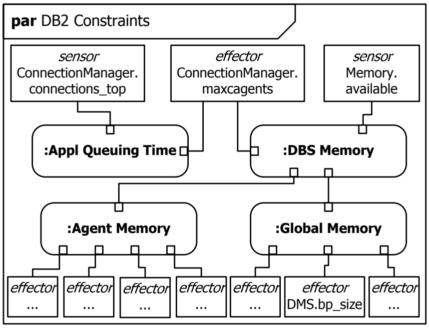


Fig. 6. Constraints Instances

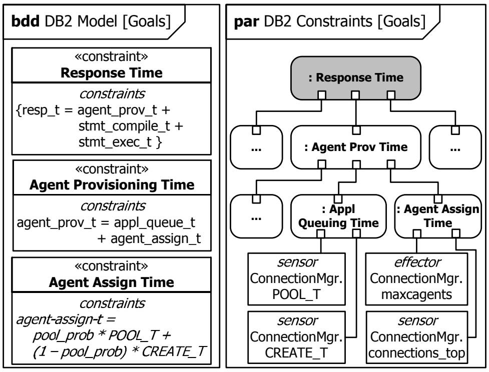

bdd DB2 Model [Goals]
«constraint» Response Time
constraints
{resp_t = agent_prov_t + stmt_compile_t + stmt_exec_t }
«constraint» Agent Provisioning Time
constraints
agent_prov_t = appl_queue_t + agent_assign_t
«constraint» Agent Assign Time
constraints
agent-assign-t = pool_prob * POOL_T + (1 − pool_prob) * CREATE_T
par DB2 Constraints [Goals]
: Response Time
...
: Agent Prov Time
...
...
: Appl Queuing Time
: Agent Assign Time
sensor ConnectionMgr. POOL_T
effector ConnectionMgr. maxcagents
sensor ConnectionMgr. CREATE_T
sensor ConnectionMgr. connections_top

Fig. 7. Objective Functions Example

Fig. 5 depicts two examples for the quantification of the behaviour of DBMS components: The *Hitratio Prediction* quantifies the expected hitratio for bufferpools depending on their size. The average time applications have the wait for an agent is approximated by *Application Queuing Time*.

To evaluate the knowledge about dependencies and component behaviour in a DBMS, the parameters of the constraints have to be mapped to the corresponding sensors and effectors. This task is the subject of the SysML parametric diagram shown in Fig. 6. A self-management logic evaluating this model would find that the expected application queuing time could be reduced by increasing the number of agents (*maxcagents*). However, as this effector is also subject to the *DBS Memory* constraint, the logic would have to find a trade-off.

4 Objective Functions

The system model described in Section 3 serves as a knowledge base for the self-management logic. It predicts the behaviour of the system under different DBS configurations. However, in order to decide whether or not the goals defined by a DBA will be met, the description of the system behaviour must be related to the goal definitions. Hence, we extend the system model by *objective functions*. These are modelled as additional constraints and represent the business goals response time, throughput, resource usage, availability and operation cost. Each of these objective functions must quantitatively describe how its value depends on the DBS configuration. The block definition diagram in Fig. 7 illustrates an example for a *Response Time* constraint, which is quantified as the sum of the waiting time for an agent (*agent_prov_t*), the time required to determine an execution plan (*stmt_compile_t*) and the time for executing the plan (*stmt_exec_t*). In order to be able to actually predict the goal values, these objective functions must be refined until they depend only on sensor and effector values. For this purpose it is possible to either re-use the component behaviour descriptions, or add new constraints. An example for objective function refinement is described for the response time goal in Fig. 7: In the block definition diagram we first define two additional constraints *Agent Provisioning Time* and *Agent Assign Time* (time for assigning an agent depending on whether or not an agent is available in the pool). These constraints are then used in the parametric diagram in a hierarchy of constraint instances, which define how the response time depends on a set of sensor and effector values. As shown in this example, the objective function constitutes the top of the constraint hierarchy, whereas the leaf nodes must be sensors and effectors described in the system model.

With the objective functions being defined in the system model, the task of self-management is to find an optimal set of effector values such that the function values are minimized (e.g. response time, operation cost) or maximized (throughput, availability). So the challenge of self-management is to find an optimal set of effector values for several, possibly opposing, objective functions. Generally speaking, the self-management has to determine an optimal configuration $\boldsymbol{x}$ for an objective vector $F(\boldsymbol{x}) = (f_1(\boldsymbol{x}), \ldots f_k(\boldsymbol{x}))$, where $f_1(\boldsymbol{x})$ to $f_k(\boldsymbol{x})$ define objective functions. In addition, it has to take into account the constraints on effector values. Finding a solution to this type of problem is the subject of multi-objective optimization (e.g. [4]).

5 Related Work

Research in the area of self-managing database systems focuses on the automation of individual administration tasks like memory management ([5]) or on-line index selection [6]. These autonomic functions do not consider relationships to other autonomic managers, side-effects, or high-level goals. Recently, also works on meeting response-time goals for multiple service classes in DBS have been published ([7], [8]). These approaches strictly focus on admission control for queries, and do not consider the adaptation of the DBS configuration.

Models with quantitative descriptions of the managed resource behaviour by now have only been used for bufferpool management in DBS. These models are used to predict the bufferpool hit ratio ([9], [10]) depending on the bufferpool size, thus allowing the definition of target service times for individual page requests. Brown et al. extend their bufferpool hit ratio prediction model in [11] by making the simplified assumption that there is a linear dependency between the hit ratio and the overall DBS response time. Unlike our approach, all quantitative models for DBS strictly focus on the bufferpool component, and do take into account the influence of the other DBS components and configuration parameters on the overall DBS perfomance.

The Common Information Model [12] provides standardized management of IT systems, independent of the manufacturer and technology. Among others, CIM defines an abstract model for database systems. However, the CIM database system model only describes general information about the DBS, like the instance name, version, the responsible DBA, and the current values of configuration parameters. The internal structure of the DBMS and a quantitative description of the system behaviour are not part of the model.

The IBM Autonomic Computing Toolkit (ACT) [13] stores information about the resource managed by an autonomic manager in a resource model. This resource model uses CIM classes to define the properties of resources, and stores additional information like check cycles, thresholds, and dependencies. However, reconfigurations cannot be automatically derived from the model, but a decision tree script must be provided, which implements this knowledge.

In order to realize self-management of complex IT infrastructures, rule-based frameworks like Accord [14] and iManage [15] have been developed. Like the ACT, these frameworks require the system administrator to define a set of actions and the conditions under which they are fired. The same approach is taken in policy management frameworks like [16]. So the decision about which reconfigurations have to be performed in order to meet business goals is not derived from a quantitative behaviour description. To overcome this limitation the Accord framework has been extended with a Limited-Look-Ahead-Controller [17]. However, the controller is limited to the optimization of a single objective function, and does not allow the creation and refinement of a system model.

In the area of web service composition there has been research on the usage of multi-objective optimization in order to meet SLAs, e.g. in [18] and [19]. Depending on the QoS requirements, multiple concrete web services are composed to realize an abstract business process. Compared to databases, the objective functions for web service composition are rather simple and the configuration alternatives are limited. However, the results show that using multi-objective optimization for meeting business goals is a feasible approach, and therefore encourage our research in applying these techniques to databases.

6 Conclusions

Currently the knowledge about the sensors and effectors in a DBS, the rules that apply to their values, and their expected effect on the system behaviour is either documented in manuals or the experience of the DBA. Representing this knowledge in a system model allows the creation of a system-wide self-management logic, which can consider the dependencies between reconfiguration actions. In addition, the quantitative description of the system behaviour can be used to ensure that business goals are met. As creating an exact quantitative model of today's complex DBMSs is a difficult task, we have proposed a graphical modelling approach, which allows the step-wise refinement of a coarse grained system model. In the future we are going to realize various coarse-grained system models for different DBMSs and evaluate the accuracy of their behaviour predictions. By comparing the different system models, we are going to identify similarities and common concepts, which will allow us the proposal of a domain-specific modelling language for DBS system models.

References

1. Weikum, G., et al.: Self-tuning Database Technology and Information Services: from Wishful Thinking to Viable Engineering. In: Bernstein, P.A., et al. (eds.) Proc. of the 28th Intl. Conf. on Very Large Data Bases, pp. 20–31. Morgan Kaufmann, San Francisco (2002)
2. Weilkiens, T.: Systems Engineering with SysML/UML, 1st edn. Morgan Kaufmann, San Francisco (2008)
3. Object Management Group: Systems Modeling Language. 1.1 edn. (2008)
4. Coello, C., et al.: Evolutionary Algorithms for Solving Multi-Objective Problems, 2nd edn. Springer, Heidelberg (2007)
5. Storm, A.J., et al.: Adaptive Self-Tuning Memory in DB2. In: Dayal, U., et al. (eds.) Proc. of the 32nd Intl. Conf. on Very Large Data Bases, pp. 1081–1092. ACM Press, New York (2006)
6. Bruno, N., Chaudhuri, S.: An Online Approach to Physical Design Tuning. In: Proc. of the 23rd Intl. Conf. on Data Engineering, pp. 826–835. IEEE Computer Society Press, Los Alamitos (2007)
7. Krompass, S., et al.: Quality of Service-enabled Management of Database Workloads. IEEE Data Eng. Bull. 31(1), 20–27 (2008)
8. Niu, B., et al.: Workload adaptation in autonomic DBMSs. In: Erdogmus, H., et al. (eds.) Proc. of the, Conf. of the Center for Advanced Studies on Collaborative Research, p. 13. IBM Press (2006)
9. Tran, D.N., et al.: A new approach to dynamic self-tuning of database buffers. ACM Transactions on Storage 4(1), 1–25 (2008)
10. Chung, J.Y., et al.: Goal-oriented dynamic buffer pool management for database systems. In: Proc. of the 1st Intl. Conf. on Engineering of Complex Systems, pp. 191–198. IEEE Computer Society Press, Los Alamitos (1995)
11. Brown, K.P., et al.: Goal-Oriented Buffer Management Revisited. In: Jagadish, H.V., Mumick, I.S. (eds.) Proc. of the ACM SIGMOD Intl. Conf. on Management of Data, pp. 353–364. ACM Press, New York (1996)
12. Distributed Management Task Force: Common Information Model (CIM) Infrastructure. 2.5.0a edn, Specification (2008)
13. IBM Corporation: A Practical Guide to the IBM Autonomic Computing Toolkit. 1st edn., Redbook (2004)
14. Liu, H., Parashar, M.: Accord: a programming framework for autonomic applications. IEEE Trans. on Systems, Man, and Cybernetics 36(3), 341–352 (2006)
15. Kumar, V., et al.: iManage: Policy-Driven Self-management for Enterprise-Scale Systems. In: Cerqueira, R., Campbell, R.H. (eds.) Middleware 2007. LNCS, vol. 4834, pp. 287–307. Springer, Heidelberg (2007)
16. Bhide, M.: et al.: Policy Framework for Autonomic Data Management. In: Proc. of the 1st Intl. Conf. on Autonomic Computing, pp. 336–337. IEEE CS Press, Los Alamitos (2004)
17. Bhat, V.: et al.: Enabling Self-Managing Applications using Model-based Online Control Strategies. In: Proc. of the 3rd Intl. Conf. on Autonomic Computing, pp. 15–24. IEEE Computer Society Press, Los Alamitos (2006)
18. Wada, H., et al.: Multiobjective Optimization of SLA-aware Service Composition. In: Proc. of the IEEE Congress on Services - Part I, pp. 368–375. IEEE CS Press, Los Alamitos (2008)
19. Chang, W.C., et al.: Optimizing Dynamic Web Service Component Composition by Using Evolutionary Algorithms. In: Skowron, A., et al. (eds.) Proc. of the IEEE/WIC/ACM Intl. Conf. on Web Intelligence, pp. 708–711. IEEE CS Press, Los Alamitos (2005)

$\mathcal{I}$-SQE: A Query Engine for Answering Range Queries over Incomplete Spatial Databases

Alfredo Cuzzocrea[1] and Andrea Nucita[2]

[1] ICAR-CNR and University of Calabria, Italy
cuzzocrea@si.deis.unical.it
[2] University of Messina, Italy
andrea@informatica.unime.it

Abstract. Spatial database systems built on top of distributed and heterogeneous spatial information sources such as conventional spatial databases underlying *Geographical Information Systems* (GIS), spatial data files and spatial information acquired or inferred from the Web, suffer from *data integration* and *topological consistency* problems. These issues make the globally-integrated spatial database *incomplete*, so that effectively and efficiently answering range queries over such databases represents a leading challenge for spatial database systems research. Inspired by these motivations, in this paper we propose $\mathcal{I}$-SQE (*Spatial Query Engine for $\mathcal{I}$ncomplete information*), an innovative query engine for answering range queries over incomplete spatial databases via meaningfully integrating *geometrical information* and *topological reasoning.* $\mathcal{I}$-SQE finally allows us to enhance the quality and the expressive power of retrieved answers by meaningfully taking advantages from the amenity of representing spatial database objects via both the geometrical and the topological level.

1 Introduction

In modern spatial database environments, data repositories collected and integrated from different spatial information sources very often coexist. Conventional spatial databases such as those underlying autonomous *Geographical Information Systems* (GIS), raw data files storing geographical information, and GIS-related Web pages are popular instances of these sources. Furthermore, the proliferation of Web- and Grid-service-based applications and systems built on top of spatial data repositories leads to an Internet-wide dissemination of spatial information sources. This phenomenon makes dealing with integration issues of spatial database systems more difficult. While the popularity of spatial data repositories within modern complex information systems, such as Web, Grid and P2P systems, is clearly an opportunity that puts the basis for further studies in the field and, symmetrically, for the industrial proliferation of spatial database systems, spatial data repositories collected and integrated from different spatial information sources also pose several research challenges. These challenges mainly concern with the presence of *incomplete information* [14,3,7] due to the

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 91–101, 2009.

heterogeneity of spatial data repositories according to several aspects, such as data models, data formats, ranges of data domains, null values handling policies, and so forth.

In heterogeneous spatial database environments like those illustrated above, *data integration* is the first issue to be faced-off [5]. Data integration has been extensively studied in the context of spatial databases (e.g., [6,10]). On the other hand, another leading challenge in spatial database systems research is represented by the issue of extending the capabilities of conventional query engines in order to make them able of dealing with the presence of several *heterogenous representations* of spatial information (e.g., [18]), which very often arise in actual GIS. This paradigm pursues the idea of representing the *same* spatial information kept in *spatial database objects* according to different levels, or *layers*, in order to enhance the expressive power of both abstraction and reasoning capabilities over spatial data. It should be noted that the latter one is a critical aspect in spatial database systems research, as modern complex information systems are more and more heterogenous in nature and kind of underlying data repositories, so that heterogeneous representations of spatial information arise accordingly. As a consequence, spatial query engines interfacing these systems have to cope with the deriving data integration issues.

Inspired by these considerations, in this paper we present $\mathcal{I}$-SQE (*Spatial Query Engine for $\mathcal{I}$ncomplete information*), a query engine for answering range queries over incomplete spatial databases, like those that derive from integrating distributed and heterogeneous spatial information sources. In particular, in the context of $\mathcal{I}$-SQE we investigate the problem of answering range queries over spatial databases where spatial information is modeled and represented according to two different levels, i.e. the *geometrical level* and the *topological level*, respectively. Also, for a sub-set of spatial database objects stored in the target spatial database interfaced by $\mathcal{I}$-SQE, one of these two levels can be missing, so that, as a consequence, incomplete spatial information occurs, and the spatial database is incomplete. The main goal of $\mathcal{I}$-SQE consists in devising intelligent techniques for answering range queries over this kind of spatial databases while overcoming incompleteness limitations.

In a conventional spatial database, spatial information is usually represented by means of detailed *geometrical properties* of spatial database objects. This because geometrical one is the most complete representation one can provide about spatial database objects. For instance, given a collection of spatial database objects, *topological relations* among these objects (e.g., containment relations) can be derived from their geometrical properties. In $\mathcal{I}$-SQE, we consider an application scenario where spatial information can be incomplete, i.e. a sub-set of spatial database objects is described by their topological relations with other spatial database objects stored in the target spatial database, whereas the geometrical information about these objects is missing. As a consequence, conventional spatial query engines, which are based on the complete availability of geometrical information about spatial database objects, are not able of effectively and efficiently answering spatial queries involving such objects.

To give an example, consider the simple case of a spatial database representing streets of a given urban area, along their geometry (i.e., geometrical information is available). Furthermore, assume that the spatial database also stores topological relations about regional areas and streets, while the geometry of regional areas is not known (i.e., topological information is available while geometrical information is not available). The simplest case of topological relation is represented by the containment one, which models the fact that a regional area A contains a set of streets $\{S_0, S_1, \ldots, S_{N-1}\}$, such that $N > 0$. If only the geometrical layer is exploited to answer range queries over the spatial database, then users only retrieve geometrical information about streets, whereas topological information on regional areas is not exploited at all. It should be noted that, in a scenario like the one described above, knowledge extracted from topological relations represents a critical "add-in" value for modern GIS applications and systems, as this information can be further exploited to enhance the knowledge discovery phase from spatial databases.

Contrary to the example above, in $\mathcal{I}$-SQE users are allowed to integrate knowledge kept in both levels, i.e. the geometrical and the topological level, respectively, thus taking advantages from both the different data representation models. Moreover, it should be noted that this paradigm is also "self-alimenting", meaning that new topological relations among queried spatial database objects can be derived by means of simple yet effective *composition rules* over already-extracted topological relations made available in the spatial database system via the query task.

The remaining part of this paper is organized as follows. In Sect. 2, we briefly review research efforts related to our research. Sect. 3 describes our technique for answering range queries over incomplete spatial databases via integrating geometrical information and topological reasoning. In Sect. 4, we present in detail $\mathcal{I}$-SQE, along its main principles, components and reference architecture. Finally, Sect. 5 discusses conclusions and future work of our research.

2 Related Work

In recent years, the proliferation of spatial data repositories has posed several challenges related to data integration issues from distributed and heterogeneous spatial information sources. For instance, the huge quantity of spatial data available on the Web leads to the possibility of their acquisition and integration within GIS, also in a semi-automatic manner [17]. Nevertheless, methods for spatial data acquisition are manyfold, and each GIS software makes use of different and heterogenous formats for representing spatial data. As a consequence, inconsistency and incompleteness arise in merged spatial data repositories, and a reasonable solution to these issues is represented by data integration techniques over such repositories.

In [13], authors propose a system for spatial data integration and sharing throughout Web services technology, via using standard Web languages and protocols such as *Geography Markup Language* (GML) [1] and *Simple Object*

Access Protocol (SOAP) [2]. In [9], a method able of evaluating queries over integrated spatial database systems is presented. Given an input spatial query Q, this method finds an optimal query execution plan for Q from the different plans computed for each feature of the integrated spatial data repository.

Models assuming the presence of *different representation layers* for spatial information have been introduced in past research efforts. In [3], a model that integrates multiple representations of *geographical maps* is presented. This model is called *Layered Spatial Data Model* (LSDM). The peculiarity of LSDM relies in the ability of representing *incomplete maps*, i.e. maps for which the geometry of the contained objects is not completely known. In addition to this, in LSDM it is also possible to represent combinatorial and topological relations among spatial database objects that qualitatively represent maps regardless of geometrical properties needed to compute them.

On the other hand, the wide availability of multi-level representation models, multi-resolution maps and spatial data mined from the Web [17] imposes us new challenges with respect to check and validation of consistency of topological relations in a spatial database system. Following this fundamental issue, [4] introduces a model for evaluating the consistency of topological relations when multi-resolution maps built on top of spatial databases are considered. As studied in [4], the main problem to be faced-off in this case relies on the fact that, in collections of multi-resolution maps that one can find in a GIS, the same spatial database object could be represented at various resolutions in different maps. This poses data as well as knowledge integration aspects to be considered.

Without doubts, topological information plays a crucial role in spatial query processing, as its semantics can be further exploited in order to improve the query capabilities of GIS integrating topology-based query engines. Nevertheless, the management of topological relations is space- and time-consuming [14]. As a consequence, devising efficient methods for representing, managing and querying topological relations plays a leading role in spatial query processing of modern GIS architectures.

In line with the considerations above, [11] proposes reducing the number of *false positives* that can be retrieved during the filtering phase of spatial selection queries via equipping each spatial database object stored in nodes of the R-tree indexing the spatial database with the so-called *Internal Rectangle* (IR). IRs are used to meaningfully infer topological relations among spatial database objects. For instance, if two IRs overlap, the actual spatial database objects overlap too. Being based on IRs, this method significantly reduces computational overheads due to computing topological relations among objects, and, as a nice consequence, the time needed to answer spatial queries involving these objects.

In [15], authors try to answer the following question: *"Which topological information on actual spatial database objects is possible to infer from topological relations among their respective MBRs?"*. They state that topological information on spatial database objects can be inferred from the *relative positions* of their respective MBRs. This fundamental insight puts the basis towards defining novel optimization strategies for efficient spatial query processing.

Finally, in [16] authors introduce a method for reducing the number of *spatial constraints* of queries via discarding those constraints that can be inferred from a sub-set of the whole (spatial) constraint set. Apart from improving the performance of spatial query evaluation, reducing the number of spatial constraints can be also useful to achieve a more compact storage representation of topological information. In a similar research initiative ([12]), *Multi-Scale Heuler Histograms* (MSHH) are proposed as a new technique for obtaining high-performance compressed representations of topological information.

In all the research initiatives reviewed above, it is always assumed that geometrical information is available for all the spatial database objects stored in the target spatial database. This allows topological relations to be computed from geometrical information in an easy manner. Contrary to this, in our research we address the relevant challenge of answering range queries over spatial databases in the presence of incomplete information, i.e. the case in which a sub-set of spatial database objects are described in the target spatial database by means of topological information only, while geometrical information associated to these objects is missing. This connotes the whole spatial information associated to these objects as incomplete.

3 Integrating Geometrical Information and Topological Reasoning for Answering Range Queries over Incomplete Spatial Databases

In this Section, we present our technique for integrating geometrical information and topological reasoning in order to answer range queries over incomplete spatial databases. This technique is implemented by algorithm `evaluateRangeQuery`, which represents the core layer of $\mathcal{I}$-SQE, our proposed query engine for incomplete spatial databases.

As highlighted in Sect. 1, in $\mathcal{I}$-SQE we focus the attention on the challenging case of dealing with incomplete spatial databases where geometrical information associated to a sub-set of spatial database objects stored in the target spatial database is missing, whereas a topological layer describing topological relations among these objects is available. Let us denote as $\mathcal{D}$ the spatial database. We denote as $G_{\mathcal{D}}$ the set of spatial database objects for which geometrical information is available, and as $T_{\mathcal{D}}$ the set of spatial database objects for which only topological information is available. For the sake of simplicity, we name as *geometrical objects* spatial database objects belonging to $G_{\mathcal{D}}$, whereas as *topological objects* spatial database objects belonging to $T_{\mathcal{D}}$, respectively. Also, we assume that spatial database objects are indexed by means of classical MBRs embedded in a high-performance R-tree indexing data structure, and that input queries are modeled in terms of two-dimensional range (spatial) queries. For instance, a typical spatial query Q belonging to this class of queries could ask if a certain spatial object O is contained by or intersects the range R of Q.

In our reference spatial database scenario, topological information is stored in the target spatial database by means of a simple yet effective two-dimensional

Table 1. A 3 × 3-array storing topological information on the spatial database objects O_i, O_j, and O_k

	O_i	O_j	O_k
O_i	E	I	O
O_j	Ct	E	I
O_k	Ct	Ct	E

array such that each entry $\langle O_i, O_j \rangle$ contains the topological relation $T_{i,j}$ among the spatial database objects O_i and O_j, i.e. $O_i \cdot T_{i,j} \cdot O_j$. Table 1 shows an example of a 3 × 3-array storing topological information on the spatial database objects O_i, O_j, and O_k. Here, E denotes the topological relation *Equal*, I *Inside*, O *Overlap*, and Ct *Contain*. For instance, $O_i \cdot I \cdot O_j$ models the topological relation stating that the spatial database object O_i is inside the spatial database object O_j (i.e., O_j contains O_i); $O_i \cdot O \cdot O_k$ means that O_i overlaps O_k, and so forth.

In a conventional spatial database system, topological information can become very large, due to the fact that a huge number of topological relations among spatial database objects can exist, as the number of topological relations is quadratic in the number of spatial database objects. As a consequence, similarly to proper spatial database objects that are indexed via high-performance R-trees, topological relations are indexed via conventional B-trees that are suitable to categorical data, and also embed efficient search algorithms for retrieving the desired information. Therefore, in our reference spatial database scenario we assume that a B-tree indexing topological relations is available.

Given a range query Q over $\mathcal{D}$ involving a set of spatial database objects belonging to $G_{\mathcal{D}} \bigcup T_{\mathcal{D}}$, our goal is to integrate geometrical information and topological information in order to provide an answer to Q, denoted by $A(Q)$. $A(Q)$ is composed by two kinds of objects: (i) geometrical objects in $G_{\mathcal{D}}$ involved by Q, for which topological relations among the geometry of these objects and the range R of Q can be easily computed; (ii) topological objects modeling topological relations between topological objects in $T_{\mathcal{D}}$ involved by Q and the range R of Q, which, contrary to the previous case, must be inferred via the method we propose in this research (recall that, for spatial database objects referred by $T_{\mathcal{D}}$, geometrical information is not available). In more detail, answering Q over $\mathcal{D}$ is performed according to a double-step approach. First, geometrical objects involved by Q are retrieved via the R-tree indexing data structure. Then, topological objects involved by Q are retrieved by means of *compositions* of topological relations between topological objects in $T_{\mathcal{D}}$ and the range R of Q. During this step, the B-tree is exploited to efficiency purposes.

Handling topological information represents a non-trivial engagement. In fact, it should be noted that topological relations retrieved during the evaluation of an input range query Q could be modeled in terms of a *disjunction of* (*basic*) *topological relations*. For instance, given a topological object O and the range R of Q, a possible disjointed expression could be: $O \cdot (Overlap \vee Inside) \cdot R$, which models the fact that O can alternatively overlap R or being contained by R. Hence, we classify the topological objects retrieved by evaluating Q into two

```
**Input:** The incomplete spatial database $\mathcal{D}$; the range query $Q$.
**Output:** The answer to $Q$, $A(Q)$.
**Method:** Perform the following steps:
$A(Q) \leftarrow \langle \emptyset, \emptyset \rangle$;
$\mathcal{A} \leftarrow \mathit{initializeArray}()$;
$\mathcal{G} \leftarrow \mathit{retrieveGeometricalObjects}(\mathcal{D},Q)$;
$\mathcal{A}.\mathit{add}(\mathcal{G})$;
**for each** $g$ **in** $\mathcal{G}\{$
  $\mathcal{R} \leftarrow \mathit{getTopologicalRelations}(\mathcal{D},\mathcal{G},g)$;
  $\mathcal{T} \leftarrow \mathit{retrieveTopologicalObjects}(\mathcal{D},\mathcal{R})$;
  $\mathcal{A}.\mathit{add}(\mathcal{T})$;
$\}$
$\mathcal{A}.\mathit{add}(Q)$;
$\mathcal{R} \leftarrow \mathit{computeTopologicalRelations}(\mathcal{A})$;
$\mathcal{T} \leftarrow \mathit{retrieveTopologicalObjects}(Q,\mathcal{R})$;
$A(Q) \leftarrow \langle \mathcal{G}, \mathcal{T} \rangle$;
**return** $A(Q)$;
```

Fig. 1. Algorithm `evaluateRangeQuery`

possible classes, namely *certain topological objects*, for which topological relations with the range R of Q can be determined exactly, and *uncertain topological objects*, for which topological relations with the range R of Q are described by a disjunction of basic topological relations, i.e. an exact representation cannot be retrieved.

In light of this, building compositions of topological relations in order to retrieve topological objects in $A(Q)$ can be questioning, due to the presence of incompleteness and uncertainty in spatial data. However, some intuitive optimizations can be devised, in order to tame computational overheads introduced by this task during the evaluation of Q. In fact, among all topological objects in $T_{\mathcal{D}}$, those that can be exploited to model compositions to be retrieved with $A(Q)$ are those for which a topological relation different from $Disjoint$ and $Universal$ with *at least* one geometrical object in $G_{\mathcal{D}}$ involved by Q exists. Recall that, given two spatial database objects O_i and O_j, $O_i \cdot Disjoint \cdot O_j$ models the fact that O_i and O_j do not have any spatial point in common (i.e., $O_i \bigcap O_j = \emptyset$), whereas $O_i \cdot Universal \cdot O_j$ models the fact that every topological relation between O_i and O_j can exist, i.e. information about the topological relation between O_i and O_j is null.

Algorithm `evaluateRangeQuery` (see Fig. 1) implements our proposed technique for answering range queries over incomplete spatial databases via integrating geometrical information and topological reasoning. Recall that, in our reference spatial database scenario, we assume that topological information about spatial database objects stored in the target spatial database is already computed and made available. `evaluateRangeQuery` takes as input an incomplete spatial database $\mathcal{D}$ and a range query Q over $\mathcal{D}$, and returns as output the answer to Q, $A(Q)$. In more detail, `evaluateRangeQuery` makes use of the following procedures: (i) `initializeArray`, which initializes the two-dimensional array $\mathcal{A}$ used as a temporary data structure to store topological information in the vest of intermediate results for the answer to Q, $A(Q)$; (ii) `retrieveGeometricalObjects`, which takes as input a spatial database $\mathcal{D}$ and a range query Q over $\mathcal{D}$, and returns as output the set of geometrical objects in $\mathcal{D}$ having a non-null intersection with Q; (iii) `add`, which takes as input a set of geometrical objects $\mathcal{G}$ and, applied to a two-dimensional array $\mathcal{A}$, adds to $\mathcal{A}$ appropriate identifiers of objects in $\mathcal{G}$; (iv) `getTopologicalRelations`, which takes as input a spatial database $\mathcal{D}$, a set of geometrical objects $\mathcal{G}$ and a geometrical object g, and returns as output the set of topological relations between g and geometrical objects in $\mathcal{G}$; (v) `retrieveTopologicalObjects`, which takes as input a spatial database $\mathcal{D}$ and a set of topological relations $\mathcal{R}$, and returns as output the set of topological objects in $\mathcal{D}$ having a topological relation different from $Disjoint$ and $Universal$ with topological objects described by $\mathcal{R}$; (vi) `computeTopologicalRelations`, which takes as input a set of topological objects (those objects whose identifiers are stored in $\mathcal{A}$), and makes use of method [8] to compute compositions of topological relations among these topological objects.

4 $\mathcal{I}$-SQE: Architecture and Functionalities

$\mathcal{I}$-SQE is characterized by a multi-layer architecture, which is shown in Fig. 2. Each layer of the $\mathcal{I}$-SQE architecture deals with a specific abstraction of the approach we propose for answering range queries over incomplete spatial databases via integrating geometrical information and topological reasoning.

The main components of $\mathcal{I}$-SQE are the following.

***Data Integration Module* (DIM).** This component deals with the problem of integrating spatial data coming from different and heterogeneous spatial information sources, such as conventional spatial databases, spatial data files, and spatial information acquired or inferred from the Web. As highlighted in Sect. 1, integrating spatial data/information poses several issues, such as ensuring the consistency of topological information over the globally-integrated spatial database. The final goal of DIM is that of collecting spatial data from the

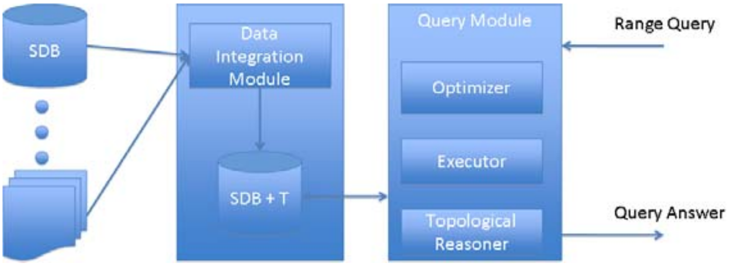


Fig. 2. $\mathcal{I}$-SQE architecture

different sources, and integrating them in a common (spatial) data format while enforcing topological consistency.

***Spatial Database with Topological Information* (SDBT).** In $\mathcal{I}$-SQE, collected and integrated spatial data repositories are materialized within a singleton spatial database where topological information is explicitly represented, named as SDBT. Storing topological information independently of geometrical one allows us to represent in the spatial database even those spatial objects for which the geometry is not known. Another important aspect to be highlighted is about the fact that, in $\mathcal{I}$-SQE, we assume that topological information about spatial database objects is maintained (and indexed) within the spatial database, while new topological relations among the range of the input query and the involved topological objects are computed on-the-fly during query evaluation.

***Query Module* (QM).** Like in a classical query engine, the component *Query Module* of $\mathcal{I}$-SQE embeds a query optimizer and a query executor, for query efficiency purposes (see Fig. 2). In addition to this, $\mathcal{I}$-SQE also embeds the component *Topological Reasoner* that is in charge of inferring topological realtions among topological objects and the range of the input query (see Fig. 3).

Summarizing, given a range query Q over an incomplete spatial database $\mathcal{D}$, $\mathcal{I}$-SQE performs the following steps in order to retrieve the answer to Q, $A(Q)$:

1. Q is parsed by the component *Query Optimizer*, which is in charge of finding an optimal query execution plan for Q, said $\mathcal{P}(Q)$;
2. Q is evaluated against $\mathcal{D}$, and a set of geometrical objects is retrieved from the integrated spatial database SDBT;
3. the component *Topological Reasoner* adds to the set of geometrical objects (retrieved according to the previous point) all the topological objects having a topological relation with the geometrical ones;

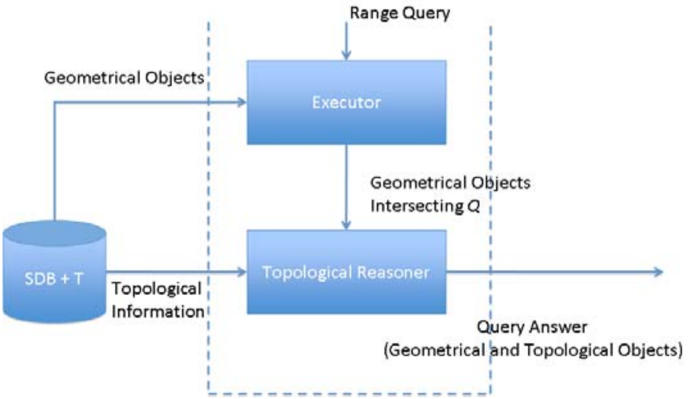


Fig. 3. The component *Topological Reasoner*

4. the answer to Q, $A(Q)$, is retrieved in the vest of a collection of geometrical and topological objects.

5 Conclusions and Future Work

In this paper, we have presented $\mathcal{I}$-SQE, a query engine for answering range queries over incomplete spatial databases via integrating geometrical information and topological reasoning. In particular, we have investigated an application scenario in which topological information exists regardless of geometrical one. We have demonstrated that, in this challenging application scenario, a conventional spatial query engine does not suffice to effectively and efficiently answer range queries, as only geometrical properties of spatial database objects are exploited in order to retrieve the final answers. Contrary to this, $\mathcal{I}$-SQE is able of enhancing the quality and the expressive power of final answers via taking advantages from both the geometrical and the topological representation of spatial database objects, thanks to a nice topological inference approach.

Future work is mainly oriented towards devising solutions for effectively and efficiently answering spatial queries over incomplete spatial databases more complex than simple range queries considered in this research, such as those embedding complex statements like join and selection-partition.

References

1. Open Geospatial Consortium, http://www.opengeospatial.org
2. World Wide Web Consortium - SOAP, http://www.w3.org/TR/soap/
3. Belussi, A., Bertino, E., Catania, B.: A Reference Framework for Integrating Multiple Representations of Geographical Maps. In: ACM GIS, pp. 33–40 (2003)
4. Belussi, A., Catania, B., Podestà, P.: Towards Topological Consistency and Similarity of Multiresolution Geographical Maps. In: ACM GIS, pp. 220–229 (2005)
5. Butenuth, M., von Gosseln, G., Tiedge, M., Heipke, C., Lipeck, U., Sester, M.: Integration of Heterogeneous Geospatial Data in a Federated Database. International Journal of Photogrammetry and Remote Sensing 62(5), 328–346 (2007)
6. Calì, A., Lembo, D., Rosati, R.: Query Rewriting and Answering under Constraints in Data Integration Systems. In: IJCAI, pp. 16–21 (2003)
7. Dehak, S.M.R., Bloch, I., Maitre, H.: Spatial Reasoning with Incomplete Information on Relative Positioning. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(9), 1473–1484 (2005)
8. Egenhofer, M.J.: Reasoning about Binary Topological Relations. In: SSD, pp. 143–160 (1991)
9. Essid, M., Boucelma, O., Colonna, F.-M., Lassoued, Y.: Query Processing in a Geographic Mediation System. In: ACM GIS, pp. 101–108 (2004)
10. Ives, Z.G., Florescu, D., Friedman, M., Levy, A., Weld, D.S.: An Adaptive Query Execution System for Data Integration. In: ACM SIGMOD, pp. 299–310 (1999)
11. Lin, P.L., Tan, W.H.: An Efficient Method for the Retrieval of Objects by Topological Relations in Spatial Database Systems. Information Processing and Management 39(4), 543–559 (2003)

12. Lin, X., Liu, Q., Yuan, Y., Zhou, X., Lu, H.: Summarizing Level-two Topological Relations in Large Spatial Datasets. ACM Transactions on Database Systems 31(2), 584–630 (2006)
13. Ma, X., Pan, Q., Li, M.: Integration and Share of Spatial Data Based on Web Service. In: IEEE PDCAT, pp. 328–332 (2005)
14. Majkic, Z.: Plausible Query-Answering Inference in Data Integration. In: FLAIRS, pp. 753–758 (2005)
15. Papadias, D., Sellis, T., Theodoridis, Y., Egenhofer, M.J.: Topological Relations in the World of Minimum Bounding Rectangles: a Study with R-trees. In: ACM SIGMOD, pp. 92–103 (1995)
16. Rodríguez, M.A., Egenhofer, M.J., Blaser, A.D.: Query Pre-processing of Topological Constraints: Comparing a Composition-Based with Neighborhood-Based Approach. In: SSTD, pp. 362–379 (2003)
17. Schockaert, S., Smart, P.D., Abdelmoty, A.I., Jones, C.B.: Mining Topological Relations from the Web. In: IEEE FlexDBIST, pp. 652–656 (2008)
18. Sheeren, D., Mustière, S., Zucker, J.-D.: How to integrate heterogeneous spatial databases in a consistent way? In: Benczúr, A.A., Demetrovics, J., Gottlob, G. (eds.) ADBIS 2004. LNCS, vol. 3255, pp. 364–378. Springer, Heidelberg (2004)

An Agent-Mediated Collaborative Negotiation in E-Commerce: A Case Study in Travel Industry

Bala M. Balachandran, Ebrahim Alhashel, and Masoud Mohammedian

Faculty of Information Sciences and Engineering,
The University of Canberra, ACT, Australia
{bala.balachandran,ebrahim.al-hashel,
masoud.mohammedian}@canberra.edu.au

Abstract. This paper examines an agent-mediated collaborative negotiation framework for e-commerce. This paper specifically focuses on travel industry. Individual customers and travel agencies will both be able to benefit from the usage of the system, since its negotiation strategies will not depend on price only, but several attributes, such as the number of rooms, the required facilities, and so on. The key issues in automating negotiation are the negotiation protocol, the negotiation object, and the negotiation strategy. Our paper addresses these issues by discussing the development of an agent-mediated e-commerce system using the FIPA compatible agent development framework, the JADE platform. Finally we provide our conclusions and discuss possible future work.

Keywords: Multi-agent systems, Intelligent Agent, Automated Negotiation, Contract-Net Protocol, JADE, FIPA, E-Commerce, Personal Travel Assistant.

1 Introduction

Agent-mediated electronic commerce (AMEC) is rapidly emerging as a new paradigm to develop distributed and intelligent e-commerce systems. Agent-mediated e-commerce systems are built upon the foundations of agent technology with a strong emphasis on the automatic negotiation [4, 6, 8, 9, 12]. We define negotiation in e-commerce as the process by which two or more parties multilaterally bargain resources for mutual intended gain. Automated negotiation takes place when the negotiating parties are represented by software agents [2, 15].

During the last decade, there has been a growth of research activities exploring the potential of automating negotiation for E-Commerce. Lomuscio et al provide an insightful overview of the existing research efforts on negotiation and describe a classification scheme for negotiation in electronic commerce [10]. Merlat discusses the importance of multi-issue negotiation for e-commerce and demonstrates a decentralised constraint satisfaction algorithm (DCSP) as a means of multi-issue negotiation [11]. He argues that it rarely the case where a single issue characterises a product or service

Travel industry is a good candidate for agent-mediated e-commerce, due to the inherent characteristics of the travel domain. For example, hotel reservation activity is well suited for automatic negotiation. In this paper, our aim is to develop a multi-agent

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 102–110, 2009.

system capable of handling the travel constraints that tie together the services such as hotel booking, flight booking and car rental, for example, time and budget constraints. We first show how to transform the travel booking problem into a multi-agent negotiation problem. Then we present our negotiation model in terms of the negotiation protocols, agent communication and negotiation strategies. We also describe details of a prototype system that is being currently developed using the JADE [7] platform. Finally, we present our concluding remarks and discuss possible future work.

2 The Travel Agency Problem

The travel agency problem has several characteristics. It is a complex domain because the travel industry is expanding rapidly and there are too many resources to connect and interact. The travel agency application has been used as a case study by many researchers in the area of e-commerce [3]. Currently customers, looking for holiday packages, depend on their travel agent to show them what is available according to their interests and budget, or visit specific content providers on the Internet such as airlines, hotel accommodators and car rental agencies. Automating these tasks by a computer system requires a complex software system involving multi-agents which are capable of solving the problem by communicating and cooperating within themselves. Typically such automated travel agency systems aim to find a bundle of services for a customer comprising travel tickets, hotel accommodation and car rental. NDumu, Collis and Nwana discuss the challenges for the creators of a personal travel assistant and propose the use of collaborative software agents as a potential solution [12]. The key challenges involved in providing integrated personalised travel services include:

- The information discovery problem
- The communication problem
- The ontology problem
- The reasoning and co-ordination problem
- The negotiation problem.

3 An Agent-Oriented Solution to the Travel Agency Problem

A multi-agent system can be defined as a collection of autonomous agents that interact with their activities in order to solve a problem that could not be solved by an individual agent [15]. In a multi-agent system, negotiation is the process by which a group of agents communicate with one another to try to resolve conflicts and reach agreement [13]. Such automatic negotiation significantly influences electronic business transactions [4]

Figure 1 illustrates our multiservice negotiation model to solve the travel agency problem. In this model, negotiations are performed through a mediator agent that acts as a unique intermediary between the customer agent and the service provider agents. The main purpose of the personal agent is to find a bundle of services matching the customer's requirements and secondly to adapt the bundle in the case of service

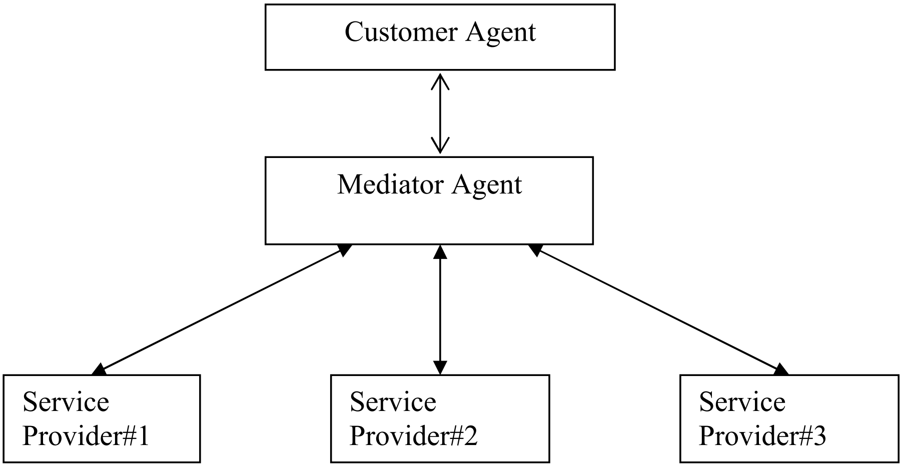


Fig. 1. Multiservice negotiation via a mediator

failure. The negotiation model we use is a bargaining game which is a process consisting of a number of offer/reply cycles, as part of an iterative improvement cycle between the mediator and participating agents. The negotiation protocol we use here is an extension of the Contract-Net Protocol [14] which adds rounds of counter proposals from the mediator agent and the other service providers.

The mediator agent requests a set of services to a set of service providers (the participants). Each participant answers either by providing a solution or by rejecting it. If the requests have been responded successfully by all the participants, then the mediator agent evaluates the package solution in terms of some ranking criteria. If the solution is not acceptable, the personal agent starts the renegotiation process with a set of modified requests. This cycle is done until a satisfactory solution is found or a predefined number of rounds is reached.

4 System Infrastructure

The architecture of our multi-agent travel booking application is made up of the following five agents: Customer Agent, Mediator Agent, Hotel Booking Agent, Flight Booking Agent, and Car Rental Agent. The behaviours of these agents are described below.

4.1 Customer Agent (CA)

The *Customer Agent* provides the user with a graphical user interface that allows the customer to enter his/her travel requirements. The Customer Agent then deals with the Mediator Agent in an attempt to obtain a suitable travel package on behalf of the customer.

4.2 Mediator Agent (MA)

The *Mediator Agent* has basically the role of travel booking. Having the offers from the participants, it tries to create an appropriate travel package that meets the user defined maximum price, thereby maximising the utility of the travel package concerning the total price. If the utility is zero it is also able to re-negotiate the prices with the participating agents. This is done in a number of user defined iterations until a travel package is found so that total price does not exceed the user defined maximum. If an appropriate travel package has been found the *Mediator Agent* books and pays for the rental car, the hotel room and the flights for the agreed offers.

4.3 Hotel Booking Agent (HBA)

The *Hotel Booking Agent* proposes hotel room offers to *the Mediator Agent.* It is also able to place reservations, if *the Mediator Agent* accepts a proposal. If a reservation has successfully placed, it also returns the reservation details.

4.4 Flight Booking Agent (FBA)

The *Flight Booking Agent* proposes flight offers to *the Mediator Agent.* It is also able to place reservations, if *the Mediator Agent* accepts a proposal. If a reservation has successfully placed, it also returns the reservation details.

4.5 Car Rental Agent (CRA)

The *Car Booking Agent* proposes car rental offers to *the Mediator Agent.* It is also able to place reservations, if *the Mediator Agent* accepts a proposal. If a reservation has successfully placed, it also returns the reservation details.

5 The Negotiation Process

Generally negotiations are characterised by their setting: bilateral, one-to-many and many-to-many. Many kinds of negotiation models exist, such as auctions, the contract-net protocol, and multi-issue negotiation. There are three key issues involved in automatic negotiation: the negotiation protocol, the negotiation strategy and the ontology. The negotiation protocol defines the messages that agents can send each other and is characterised by successive messages exchanged between an initiator and participants as in the contract net protocol framework [5]. It is necessarily open and public. The negotiation strategy, on the other hand, is the way in which a party acts within the protocol specification to get the best outcome of the negotiation. It is therefore necessarily private. The ontology is a way for categorising domain objects so that they are semantically meaningful to a software agent.

In our model, each of the service provider agents follows a certain price strategy for hotel room, rental car and flight offers, when re-negotiation of prices with *the*

Mediator Agent becomes necessary. The price strategies are thereby the reduction of the offer price. Some example strategies are as follows:

- After a certain number of iterations by a certain percentage (e.g. after 2 iterations by 20%);
- After a certain number of iterations by a fixed amount (e.g. after 3 iterations by 100 Dollars) or
- After each iteration by a certain percentage (e.g. in the first two iterations by 10%).

The Mediator Agent has the broker role. In this role it forwards the requests from the *Customer Agent* to the *appropriate service provider agent* and returns the results back to the *Customer Agent*. It thereby acts as a mediator that understands the interaction protocol the *Customer Agent* is speaking and translates it into an interaction protocol that the *Service Provider Agent* understands.

We consider four different scenarios in the negotiation process:

Successful booking without iterations: In this scenario the booking is successful. This means there are flight-, car- and hotel-room offers for the given period and a travel package that does not exceed the given maximum price can be built.

Unsuccessful booking without iterations: This scenario shows a non-succeeding booking.

The scenario can be divided into two cases. In the first case the booking fails because no offers (hotel, car or flight) are available for the trips target period. In the second case one or more offers are available; however no travel package can be built for the given maximum price.

Unsuccessful iterated booking with iterations: This scenario shows a non-succeeding booking process that includes a number of iterations. During the iterations the *Service Provider Agents* can reduce the price of their offers. On the other hand the *Mediator Agent* tries to maximize the value of the utility function discussed below. The booking process fails if the *Mediator Agent* is not able to build a travel package that does not exceed the given maximum price, or if no offers (hotel, car or flight) are available for the required period.

Successful iterated booking with iterations: The last scenario covers a successful booking process after a number of iterations. The Mediator Agent initiates a negotiation process with the booking agents in order reach a deal that is suitable for the customer. In this case the booking agents reduce the price of their offers according to their own strategy.

5.1 Agent Communication

The agents described for the proposed multi-agent travel booking system use the FIPA Iterated Contract Net Interaction Protocol for the communication and

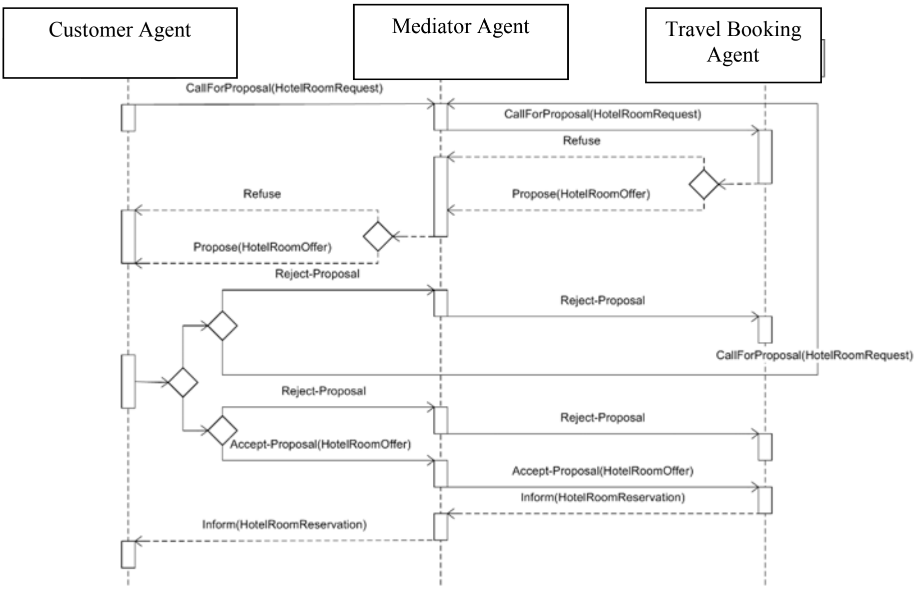

Customer Agent
Mediator Agent
Travel Booking Agent
CallForProposal(HotelRoomRequest)
CallForProposal(HotelRoomRequest)
Refuse
Propose(HotelRoomOffer)
Refuse
Propose(HotelRoomOffer)
Reject-Proposal
Reject-Proposal
CallForProposal(HotelRoomRequest)
Reject-Proposal
Reject-Proposal
Accept-Proposal(HotelRoomOffer)
Accept-Proposal(HotelRoomOffer)
Inform(HotelRoomReservation)
Inform(HotelRoomReservation)

Fig. 2. Communication between the agents during the travel booking

negotiation [5]. Figure 2 illustrates the communication between the Customer Agent, the Mediator Agent and the TravelBookingAgent via the FIPA Iterated Contract Net Interaction Protocol.

The protocol starts with a Call- For Proposal message. The message contains either a CarRentalRequest, a FlightTicketRequest or a HotelRoomRequest depending on the roles offered by the TravelBookingAgent, The request and offer types send between the agents are defined in a travel booking ontology. If the TravelBookingAgent can propose an offer, it sends either a CarRentalOffer, a FlightTicketOffer or a HotelRoomOffer to the MediatorAgent. The MediatorAgent then evaluate all received offers. If the price of a travel package is lower or equal than the budget, the MediatorAgent accepts the offers of the appropriate travel booking agents. If all round prices of all possible travel packages are greater than the budget, the Mediator initialises further negotiation iterations by sending CallForProposal messages to all travel booking agents.

During the iterated negotiation the Mediator Agent tries to build a travel package whose value of the all round price utility is greater than 0. See Figure 3 below. This means that the all round price is between 0 and the maximum value for the price. If more than one combination of proposals can fulfill this requirement, the Mediator Agent tries to maximize the value of the utility function for selection of one of them. In other words always the cheapest travel package will be selected.

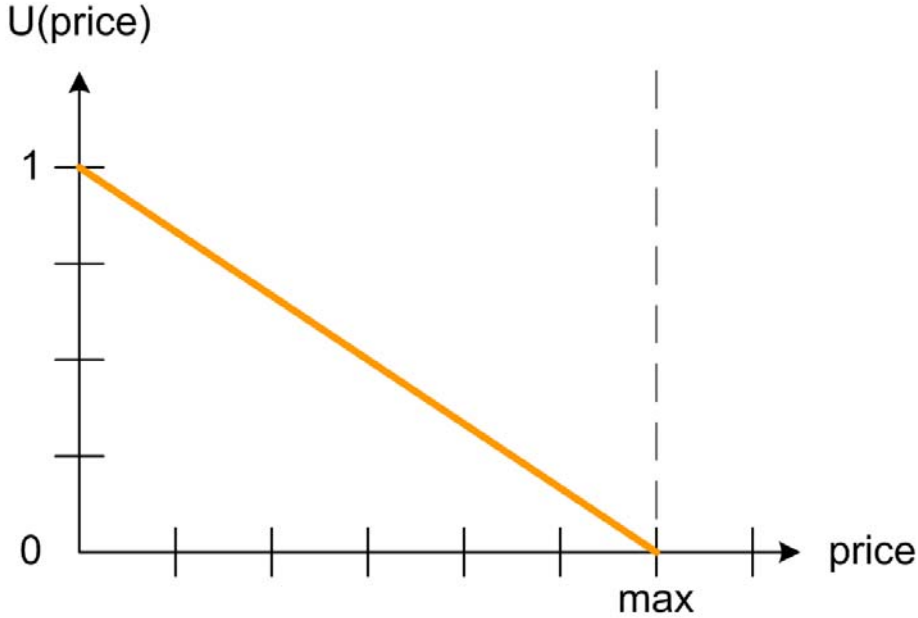


Fig. 3. Price Utility Function

6 Developing Multi-agent Systems with JADE

Using Gaia methodology, analysis and design models have been created for the travel industry support system [1]. The system implementation is being carried out in JADE environment using the Gaia models. JADE has been selected because of its open-source, ease of use and compliant with the FIPA specifications [5]. Agent communication is probably the most fundamental feature of JADE and is implemented accordance with the FIPA specifications. The JADE communication paradigm is based on asynchronous message passing. Each agent is equipped with an incoming message box and message polling can be blocking or non-blocking. FIPA specifies a set of standard interaction protocols such as FIPA-request, FIPA-query, and so on. A message in JADE is implemented as an object of the jade.lang.acl.ACLMessage object and then calling the send () method of the Agent class. We have implemented partially the travel industry support system in JADE 3.5 platform and our implementation process is going on.

The major components of our prototype system are illustrated in Figure 4.

User Interface: The customer applies the UI for accessing the travel support system such as inserting the search criteria, choosing and booking the appropriate travel package, as well as receiving the search and booking results, the system messages and confirmations.

Database: It contains travel information including flights, accommodation and cars, as well the customer details.

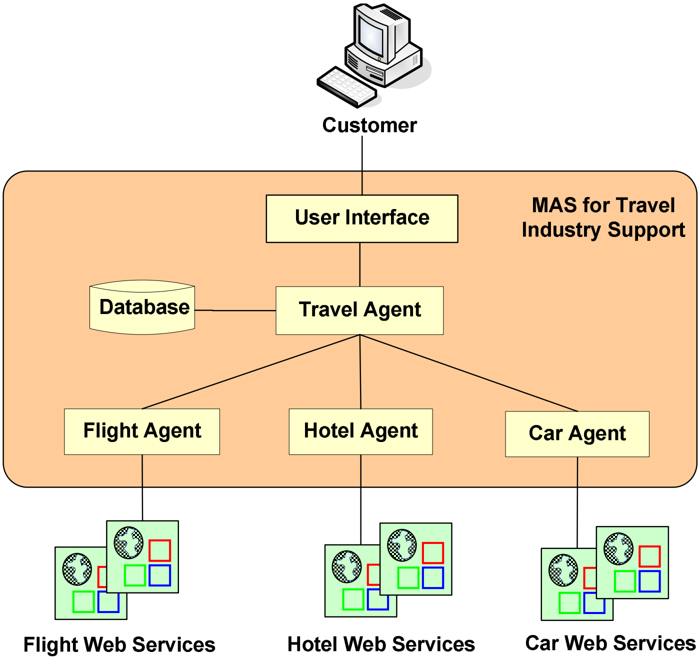


Fig. 4. The components of the prototype travel booking system

Agents: The system has five types of agents to approach the goal for finding and booking the appropriate travel packages as described in section 4. Mediator agent is responsible for finding the travel package using the given search criteria by the customer and incorporates with the other agents to find out the booking availability and book the travel package. Flight, Hotel and Car agents interact with the web services to retrieve the information of availability for booking and to create booking. These agents are activated when Mediator agent requests the action.

7 Conclusions

Developing a multi-agent e-commerce system involves many challenges, including agent coordination, agent negotiation, agent communication, system infrastructure, intelligence of agents and system implementation. In this paper, we presented some of the problems inherent to automating integrated travel users through the use of software agents. We presented a collaborative agent-based approach that makes personal travel assistant development possible. We described an agent-mediated coordination and negotiation to solve the problem. The customer's requirements are met through a series of negotiations between the mediator agent and the service providers. Each agent is capable of using its own strategy to handle the requests. This work has demonstrated that agent technology is a very promising tool to start address real business problems. Even though we have implemented some parts of the multi-agent travel booking application, we still need to fully implement the designed features for travel booking in our future work.

References

1. Balachandran, M.B., Enkhsaikhan, M.: Development of a Multi-agent system for Travel Industry Support (CIMCA 2006 and IAWTIC 2006), Sydney, Australia (2006)
2. Beam, C., Segev, A.: Automated Negotiations: A Survey of the State of the Art. Wirtschaftsinformatik 39(3) (1997)
3. Nwana, H.S., et al.: Agent-Mediated Electronic Commerce: Issues. In: Challenges and Some Viewpoints, Autonomous Agents 1998, MN, USA (1998)
4. Fasli, M.: Agent Technology for eCommerce. John Wiley and Sons, UK (2007)
5. FIPA: The Foundation for Intelligent Physical Agents. FIPA Iterated Contract Net Interaction Protocol, `http://www.fipa.org/specs/fipa00030/`
6. Guttman, R., Moukas, A., Maes, P.: Agent-Mediated Electronic Commerce: A Survey. Knowledge Engineering Review 13(2), 147–159 (1998)
7. JADE (2006), Java Agent Development Environment, `http://jade.tilab.com/`
8. Lin, R.J., Cho, S.-C.T.: Mediating a Bilateral Multi-Issue Negotiation. Electronic Commerce Research and Applications 3(2) (2004); Proc. of CEC 2003
9. Liu, K., Feng, Y.: E-Commerce Oriented Automated negotiation Based on FIPA Interaction Protocol Specification. In: Proceedings of the Sixth International Conference on machine Learning and Cybernetics, Hong Kong, August 19-22 (2007)
10. Lomuscio, A.R., Wooldridge, M., Jennings, N.R.: A Classification Scheme for Negotiation in Electronic Commerce. In: Group Decision and Negotiation, vol. 12, pp. 31–56. Kluwer Academic Publishers, Dordrecht (2003)
11. Merlat, W.: An Agent-Based Multiservice Negotiation for ECommerce. BT technical Journal 17(4), 168–175 (1999)
12. Ndumu, D.T., Collis, J.C., Nwana, H.S.: Towards desktop personal travel agents. BT Technology J. 16(3) (July1998)
13. Rosenschein, J., Zlotkin, G.: Rules of Encounter: Designing Conventions for Automated Negotiation among Computers. MIT Press, Cambridge (1994)
14. Sandholm, T.: An Implementation of the Contract Net Protocol Based on Marginal Cost Calculations. In: Proceedings of the Eleventh National Conference on Artificial Intelligence (AAAI 1993), Washington, D.C. (1993)
15. Wooldridge, M.: Introduction to Multiagent Systems, 2nd edn. John Wiley and Sons, UK (2009)

The Role of Ontology in Modelling Autonomous Agent-Based Systems

Ebrahim Alhashel, Bala M. Balachandran, and Dharmendra Sharma

School of Information Sciences and Engineering
University of Canberra, ACT 2601, Australia
{ebrahim.al.hashel,bala.balachandran,
dharmendra.sharma}@canberra.edu.au

Abstract. An agent-based system is characterised by an agent's autonomous behaviour, which behaviour is the main difference between the concepts of agent and object. Agent autonomous behaviour is the ability of an agent to cooperate instead of integrate; therefore, the structure of agent-based systems consists of loosely coupled agents. In such an environment, the relationship between the agents is unlocked, so conventional, predefined integration software techniques are not an option because the agents need an open-architecture type of integration (cooperation) to achieve their tasks jointly. The aim of this research paper is to provide an evidence of how the ontology approach can play a role in modelling agent autonomous behaviour. The research explores the ontology software technologies used for semantic web applications, and designs a case study as an example of a set of services. In the implementation phase, the research uses the web ontology software development languages XML, RDFS, OWL, and Altova semanticWork to set up and develop the case study. The result is presented and plans for future work are discussed.

Keywords: Software engineering, knowledge-based system, agent-oriented software development, AOSE, Ontology, semantic Web, Web Ontology Language.

1 Introduction

Agent-based computing offers a new software design for managing the inherent complexity of software systems [1]. Unlike an object-based system or any other software paradigm, an agent-based system is characterised by agent autonomous behaviour wherein the agent can act without predefined rigid manipulation or links, such as object invocation and primary and foreign key techniques. To design agent autonomous behaviours, there must be an open integration architecture that permits an agent to function in its environment and to cooperate with other agents. The theory of the autonomous agent specifies that the agents must have the ability to act in a group toward solving a common problem—in other words, to act cooperatively [2].

This paper presents the concepts of Ontologies and the existing ontology-based software development tools to increase understanding of how efficiently this technology can be utilised for developing agent-based paradigms, particularly in

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 111–118, 2009.

modelling agent autonomous behaviour. Given the view that agent behaviour requires an open-architecture mechanism, we consider the ontology techniques to be a powerful tool to represent agent autonomous concepts. The idea is that, *if the agent can realize the meaning of things (objects), can match its own capability with these objects' meanings, and then can identify its domain objective, this is an efficient technique to support the agent autonomy.*

To illustrate the research vision, we develop a travel agency system (TAS) e-commerce case study defined in [3], and then demonstrate the implementation processes. The architecture of TAS is divided to two main components: the services and the agents. The services are the subsystems, for example, Car, Hotel, and Flight. The services must be ontologically defined and then made readable and understandable for the system's agents. The agents must be able to discover the services' functionalities (what the services do exactly). The agents are designed in two categories: professional agents and skill agents. The professional agents are responsible for the main user goal (what the user intends the system to do), for example, travel to Melbourne. The skill agents are responsible for representing a group of services with the same functionalities, for example, a set of services providing cars for hire.

The ontology of the system's services provides an efficient search based on the functionality (what the service does), rather than only on the service name (key word). The interpretation of the inter-operations of the services is the main element required to create automatic configuration between the services because this interpretation process allows one service to use other services' facilities to satisfy the user's goal. For example, a holiday package needs integration among three agent services: hotel booking, flight reservation, and car hire.

This paper is organised as follows. Section 2 discusses the agent autonomous architectures. Section 3 briefly demonstrates the development of the ontology using the existing software development tools. Section 4 uses a TAS case study to illustrate how an agent-based system can be developed using ontology technologies. Section 5 summarises our conclusions and relates plans for future work.

2 Autonomous Agent Architectures

Many agent-based software architectures have been proposed, with each focusing on a specific agent approach or designed toward a specific applications domain. This multiplicity of approaches has resulted in many different architectures. However, agent autonomous behaviour has not yet been implemented or designed under any architecture.

The agent-based software is a new paradigm for which some important questions remain, including which components constitute agent-based systems, to what level the agent concept has been implemented, how these concepts are represented and how they fit together to form the required system. There is no definite answer to these questions, but the existing agent development tools are affected by this state of affairs. For example, agent software engineering development methodologies such as Gaia [4] view the system according to an organisational structure approach. MaSE [5] combines several pre-existing models into a single structure methodology and then focuses on goal hierarchy and system roles. Prometheus [6] develops BDI independent agents. Tropos [7] emphasises agent components as actors, goals, and plans. ROADMAP [8] is designed from knowledge perspectives and is normally used as an extension of Gaia [3].

Other agent-based approaches can be examined in terms of their development environments. JADE [9], for example, uses FIPA agent architecture. JADEX [10] is a Belief Desire Intention (BDI) reasoning engine that allows for programming intelligent software agents in XML and Java. JACK [11] develops BDI agent architecture [12]. Cougaar [13] is used for designing and building Society-Community agent architecture, and Aglet [14] deals with the mobility of an agent.

There are also design-objective models, such as RETSINA [15], which generates a multi-agent architecture based on peer coordination. The Zeus [16] agent development tool uses a graphical user interface (GUI) to define agent components and generate java code. A few theoretical attempts have been made in the same stream. Luck [17] discussed an alternative state architecture and, by using adoption strategy, showed how agents emerge from objects. All of these attempts have endeavoured to formulate software agent architecture using different approaches. For more details on these approaches, refer to [18].

3 Ontology in Software Engineering

The use of ontology in software engineering comes into demand in active research areas. The new trend in software engineering is to construct a logic-ware that possesses a "reason" property in its processing and, in this, ontology plays a main role and can be used to represent a meaning (knowledge) of the application domain. It has been shown that, in merging the ontology into the software, engineering ontology using UML has some drawbacks. To show the use of UML, class diagrams will be used to represent concepts and their attributes and relationships between concepts. For axiom, an Object Constraint Language (OCL) is used and attached to the concept notes, since UML lacks formal semantics [19]. [20]. However, UML does not own the full potential to model the Ontologies domain model and, for this reason, this research paper introduces its own modelling diagram to represent both agent abstractions and the ontology constraints.

Other software engineering for modelling ontology can be done by extending the database design techniques and the entity-relationship diagram (ERD) [21]. Asuncion et al. demonstrated the utilisation of ERD, and then mentioned the use of the model extension using HERM (high-order entity-relationship model) to add complex attribute types (key constraints, generalisation and specialisation relationships, etc.) [22]. However, the main drawback of using extended ERD is that heavyweight ontology cannot be modelled. A set of ontology development tools is currently available. This research focuses only on the relevant tools, specifically those based on eXtensible Markup Language (XML), such as Ontology Web Language (OWL) [23] and its extension, Resources Description Framework (RDFS). OWL provides more advance inference to answer queries not necessary predefined properly since OWL has more advanced generalisation properties than RDFS. OWL is part of the growing stack of W3C recommendations related to the Semantic Web. The following are more illustrations of the use of semantics, taken directly from [23]:-

- XML provides a surface syntax for structured documents but imposes no semantic constraints on the meaning of these documents.
- XML Schema is a language for restricting the structure of XML documents and also extends XML with datatypes.
- RDF is a datamodel for objects ("resources") and the relationships between them and provides a simple semantics for this datamodel, which can be represented in an XML syntax.
- RDF Schema is a vocabulary for describing properties and classes of RDF resources, with a semantics for generalisation-hierarchies of such properties and classes.
- OWL adds more vocabulary for describing properties and classes: among others, relationships between classes (e.g., disjointedness), cardinality (e.g., "exactly one"), equality, richer typing of properties, characteristics of properties (e.g., symmetry), and enumerated classes.

4 Case Study

Planning a travel package that involves a flight, a hotel, and a car reservation on a given date and time is a complex task to automate, particularly when more than one website is involved. Coordination between each service is the key to accomplishing a task such as this. Currently, this type of task is performed by humans after searching and accessing several websites and making decisions based on preferences. In this environment, the concept of an intelligent software agent has great potential for helping the customer get the best deal on a travel package.

The proposed **Travel Agency System (TAS)** is a multi-agent system designed to obtain travel packages for users based on their preferences. This process begins when the TAS scenario is defined to capture the goals and sub-goals. For this purpose, the extended Prometheus development methodology defined by [24] is applied. For the result, see Figures 1 and 2.

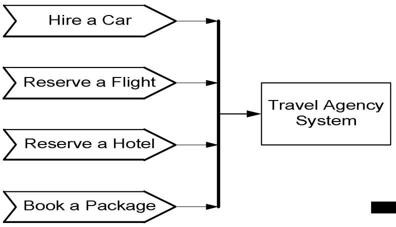


Fig. 1. System goals diagram

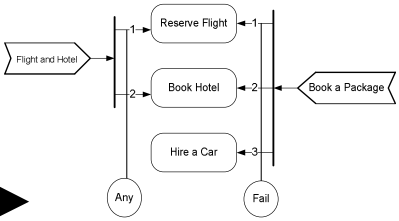


Fig. 2. Converting goals to tasks

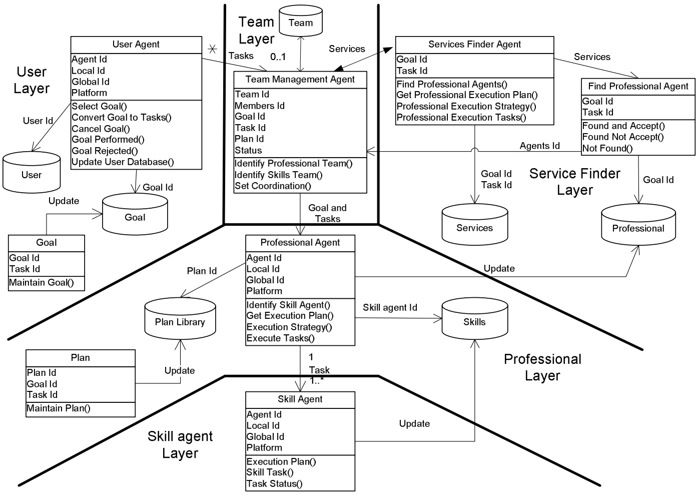


Fig. 3. TAS agent classes architecture

It is not practical to show all the implementation details here, but Figure 3 shows the TAS class diagram based on the TAS architecture. The TAS architecture is divided into five layers: User agent, Services finder, Team management, Professional agent, and Skill agent. The main design principle used in TAS architecture is that partitioning the services layer from the team management layer isolates services (what will be done) from the execution (who will do it). This partitioning is made possible because services are defined semantically using resources description framework language (RDF) in parallel with ontology web language (OWL) classes, and extensible mark-up language (XML) for data representations.

TAS User Layer: In this process, the user's goals are captured and then analysed further until all the goals and related tasks are identified.

Team Management Layer: At this layer, both the system agent and the agent components are identified, and the team database, the plan database, and the professional agents required to achieve the goals are recognised.

Services Finder Layer: The functionality model for each service should be developed with the execution sequence. (Refer to the elements of the service profile listed in [25].) Each row has to be defined at this stage, and the XML document that represents the services data that will be used by OWL has to be developed. Figure 4 provides a selected segment code for the car rental ontology.

Professional agent layer: According to the TAS scenario, the goal layer consists of three professional agents: Car, Flight, Hotel, and Package. The main elements of the agent inter-structure have to be developed for every agent (Figure 5).

Skill agent Layer: The Skill agent layer is concerned with three aspects of the process: the professional agent under which it performs, the execution sequence, and which services it executes. The plan and the database of the TAS are also considered at this stage.

```
<?xml version="1.0"?>
<rdf:RDF
 xmlns:owl="http://www.w3.org/2002/07/owl#"
 xmlns:rdf=http://www.w3.org/1999/02/22-rdf-syntax-ns#
 xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
 xmlns:rent="http://www.rent.com/ontologies/Rent#"
 xmlns:xsd="http://www.w3.org/2001/XMLSchema#">

 <rdf:Description
  rdf:about="http://www.rent.com/ontologies/Rent#Rent">
  <rdf:type>
   <rdf:Description
    rdf:about="http://www.w3.org/2002/07/owl#Class"/>
  </rdf:type>
 </rdf:Description>
 <rdf:Description
  rdf:about="http://www.rent.com/ontologies/Rent#Car">
  <rdf:type>
   <rdf:Description
    rdf:about="http://www.w3.org/2002/07/owl#Class"/>
  </rdf:type>
  <rdfs:subClassOf>
   <rdf:Description
  rdf:about="http://www.rent.com/ontologies/Rent#Rent"/>
    </rdfs:subClassOf>
 </rdf:Description>
 <rdf:Description
  rdf:about="http://www.rent.com/ontologies/Rent#Type">
  <rdf:type>
   <rdf:Description
    rdf:about="http://www.w3.org/2002/07/owl#Class"/>
    </rdf:type>
 </rdf:Description>
 <rdf:Description
  rdf:about="http://www.rent.com/ontologies/Rent#Brand">
  <rdf:type>
   <rdf:Description
    rdf:about="http://www.w3.org/2002/07/owl#Class"/>
    </rdf:type>
 </rdf:Description>
 .
 .
  </rdf:RDF>
```

Fig. 4. Selected segment of OWL codes for car rent ontology

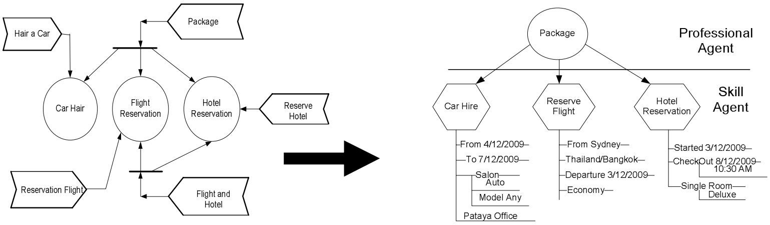


Fig. 5. Relationships between professional agent, Skill agent, and goals

5 Conclusion and Future Work

The agent-based system introduces a new feature that adds challenges to software engineering. Many tools and methodologies have been developed to engineer agent behaviour, and all have their advantages and disadvantages. Developing an ideal agent-based system agent's autonomous behaviour requires agents to cooperate instead of predefining integration; otherwise, agent becomes object.

The TAS architecture is an agent-based system designed to enable agent to exercise its autonomous behaviour. To implement agent autonomous feature, partitioning the system's characteristics from the agent's actions is essential; in TAS, the system characteristics are partitioned from the agent's acts by using a services-related semantic approach and the web ontology tools OWL, RDFS, and XML. This technique provides an efficient methodology for developing cooperative software agent-based systems. TAS is currently in the implementation phase, so the result of this project will be reported in the near future.

References

[1] Odell, J.: Objects and Agents Compared. Journal of Object Technology 1, 41–54 (2002)

[2] Alhashel, E., Mohammadain, M.: Illustration of Multi-agent Systems. Presented at CIMCA, Austria - Vienna (2008)

[3] Alhashel, E., Balachandran, B., Sharma, D.: Comparison of Three Agent-Oriented Software Development Methodologies: ROADMAP, Prometheus, and MaSE. In: Apolloni, B., Howlett, R.J., Jain, L. (eds.) KES 2007, Part III. LNCS (LNAI), vol. 4694, pp. 909–916. Springer, Heidelberg (2007)

[4] Wooldridge, M., Jennings, N., Kinny, D.: The Gaia Methodology for Agent-Oriented Analysis and Design. Presented at Autonomous Agents and Multi-Agent Systems, The Netherlands (2000)

[5] Wood, M.F., Deloach, S.A.: An Overview of the Multi-agent System Engineering Methodology. Presented at First International Workshop on Agent-Oriented Software Engineering, Limerick Ireland (2001)

[6] Padgham, L., Michael, W.: Developing Intelligent Agent Systems, vol. 1, pp. 23–82. John Wiley & Sons, Ltd, England (2004)

[7] Bresciani, P., Giorgini, P., Giunchiglia, F., et al.: Tropos: An Agent-Oriented Software Development Methodology. Presented at Autonomous Agents and Multi-Agent Systems, Netherlands (2004)

[8] Juan, T.: The ROADMAP Meta-Model for Intelligent Adaptive Multi-Agent Systems in Open Environments (2003)

[9] Bellifemine, F., Caire, G., Greenwood, D.: Developing Multi-agent Systems with JADE. Wiley, Chichester (2007)

[10] Braubach, L., Pokahr, A., Lamersdorf, W.: Jadex: A BDI-Agent System Combining Middleware and Reasoning. In: Software Agent-Based Applications, Platforms and Development Kits. Whitestein Series in Software Agent Technologies and Autonomic Computing, pp. 143–168. Birkhäuser, Basel (2005)

[11] JACK Development Toolkit (07/07/2008), http://www.agent-software.com/

[12] Kinny, D., Georgef, M., Rao, A.: A Methodology and Modelling Technique for System of BDI Agents. Presented at Modelling Autonomous Agents in a Multi-Agent World MAAMAW 1996, Germany (1996)

[13] Cognitive Agent Architecture (Cougaar) Project (17/04/2009), `http://www.cougaar.org/`

[14] Aglet Development Toolkit (23/07/2008), `http://aglets.sourceforge.net/`

[15] RETSINA Development Toolkit (24/06/2008), `http://www.ri.cmu.edu/projects/project_76.html`

[16] Development toolkit (12/09/2008), `http://labs.bt.com/projects/agents/zeus/`

[17] Luck, M., d'Inverno, M.: A Formal Framework for Agency and Autonomy. Presented at International Conference on Multi-Agent Systems (1996)

[18] Luck, M., Ashri, R., D'Inverno, M.: Agent-Based Software Development. Artec House Inc., London (2004)

[19] Siricharoen, V.W.: Ontologies and Object models in Object Oriented Software Engineering. Presented at International Multi Conference of Engineers and Computer Scientists 2007, Hong Kong, March 21-23 (2007)

[20] Gomez-Perez, A., Fernandez-Lopez, M., Corcho, O.: Ontological Engineering with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web. Springer, Heidelberg (2004)

[21] Fankam, C., Jean, S., Bellatreche, L., et al.: Extending the ANSI/SPARC architecture database with explicit data semantics: An ontology-based approach. In: Morrison, R., Balasubramaniam, D., Falkner, K. (eds.) ECSA 2008. LNCS, vol. 5292, pp. 318–321. Springer, Heidelberg (2008)

[22] Thalheim, B.: Entity-Relationship Modeling: Foundations of Database Technology. Springer, Berlin (2000)

[23] OWL Web Ontology Language (01/06/2009), `http://www.w3.org/TR/owl-features/`

[24] Alhashel, E., Balachandren, B., Sharma, D.: Enhancing Prometheus to Incorporate Agent Cooperation Process. Presented at International Conference on Computational Intelligence for Modelling, Control and Automation - CIMCA, Austria - Vienna (2008)

[25] OWL-S: Sementic Markup for Web Services (10/04/2009), `http://www.w3.org/Submission/OWL-S/`

Multi-Agent Systems in Quantum Security for Modern Wireless Networks

Xu Huang and Dharmendra Sharma

Faculty of Information Sciences and Engineering,
University of Canberra, ACT 2601, Australia
{Xu.Huang,Dharmendra.Sharma}@canberra.edu.au

Abstract. Security in wireless networks has become a major concern as the wireless networks are vulnerable to security threats than wired networks. The 802.11i wireless networks uses 4 way handshake protocol to distribute the key hierarchy in order to encrypt the data communication. In our previous research work [2], [3], we have investigated Quantum Key Distribution (QKD), for key distribution in 802.11 wireless networks. The whole communication flow of our proposed protocol can be split into several key processes. It can be seen that these processes can be implemented efficiently using Software Agents. In this paper we shall focus on the use of Software Agents in quantum cryptography based key distribution in WiFi wireless networks.

1 Introduction

Wireless communications are becoming ubiquitous in homes, offices and enterprises with its ability to provide high-speed, high-quality information exchange between portable devices.

WiFi networks uses 802.11 and 802.1X for association and authentication process. The authentication of the end users is essential in wireless networks as the wireless medium is accessible openly. A lot of research papers highlighted the security flaws of wireless networks based on 802.11 [10], [11], [12], [13]. Most of those are happening in the form of Denial of Service (DoS) attacks, Main-in-the-Middle (MiM) attacks, session hijacking (SH) etc.

In our previous [2], [3] and subsequent work, we have come up with a novel protocol to perform the key management in WiFi networks. Software agents can deliver much needed intelligent behavior to WiFi networks especially in case of adversary attacks. In this paper, we explain how Multi Agent Systems (MAS) can be used to perform the key exchange in WiFi networks.

2 Integrating Quantum Key Distribution in IEEE 802.11i Networks

The IEEE 802.11 Task Group has come up with an amendment to the IEEE 802.11 standard [4] called IEEE 802.11i [1] in 2004 to address the security flaws encountered

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 119–126, 2009.

in its initial design. IEEE 802.11i separates the authentication and encryption key management. For authentication 802.11i uses IEEE 802.1X [5], [6] and pre-shared key. IEEE 802.1X offers an effective framework for authenticating, managing keys and controlling user traffic to protect large networks.

Once this process completes, the 802.11i 4-way handshake process takes place and ensures the 802.11i key hierarchy to establish at both ends. This key hierarchy consists of several keys, namely: Pairwise Master Key (PMK), EAPOL-Key Confirmation Key (KCK), EAPOL-Key Encryption Key (KEK), Group Temporal Key (GTK) and Temporal Key (TK).

2.1 Wireless with Using of Quantum Cryptography

Though the use of quantum cryptography in wireless communications is still premature, the "unconditional security" [19] of quantum cryptography offers much needed security for wireless networks. At present lot of research work and commercial implementations are happening in this area [17], [18], [23], [24]. Several QKD protocols such as SARG04 [7], BB84 [8], B92 [9] and six-state [10] exist as of now. Out of those, BB84 has proven in practical networks. SARG04 protocol is an improved version of BB84 by eliminating Photon Number Splitting (PNS) attacks. As BB84 does, SARG04 protocol operates in two stages: Quantum channel and Classical channel. In the first stage, photon transmission takes place via quantum channel between two parties. Each of these photons represent a binary bit value of the secrete key. During the second stage, the two parties communicate with each other as per the SARG04 protocol to obtain the secrete key. The second stage comprises of four main phases: (1) Raw Key Extraction (Sifting), (2) Error Estimation, (3) Reconciliation and (4) Privacy Amplification. Further investigation of SARG04 protocol is beyond the scope of this paper.

2.2 QKD Based Key Exchange in 802.11i

Figure 1 shows the full 802.11i protocol communication including the quantum key exchange. Flows 1 to 6 illustrate the IEEE 802.11 association and authentication process. During this process, the Supplicant creates an 802.11 association with the Authenticator. Once the IEEE 802.11 association is completed, the IEEE 802.1X authentication starts with the Supplicant sending EAP-Start message to the Authenticator. This process is shown by flows 7 to 13 of Figure 2. In our work, we choose to use EAP types such as EAP-TLS, EAP-TTLS etc. that offer mutual authentication between the Supplicant and the Authenticator.

At the end of this process, i.e. flow 13 of Figure 2, both Supplicant and Authenticator are in possession of Pairwise Master Key (PMK). Then the communication switches to quantum channel and the photon transmission takes place from the Supplicant towards the Authenticator. Once the quantum transmission finishes, communication channel switches back to wireless channel. Afterwards the SARG04 protocol takes place as shown in flows 15 to 18 in Figure 2 to obtain the final secrete key. From this key, the 802.11i key hierarchy containing PTK, KCK, KEK, TK and GTK can be retrieved. The TK is used to encrypt data for the subsequent data communication. This whole process is not explained in detail in this paper as our focus is on the use of agents for this protocol.

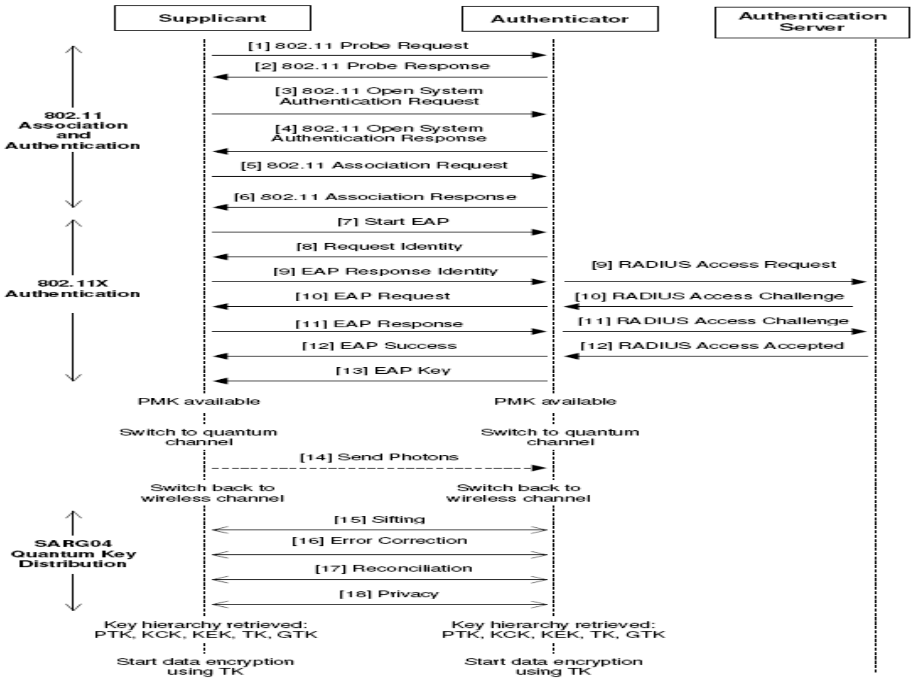

Supplicant
Authenticator
Authentication Server
802.11 Association and Authentication
[1] 802.11 Probe Request
[2] 802.11 Probe Response
[3] 802.11 Open System Authentication Request
[4] 802.11 Open System Authentication Response
[5] 802.11 Association Request
[6] 802.11 Association Response
802.11X Authentication
[7] Start EAP
[8] Request Identity
[9] EAP Response Identity
[9] RADIUS Access Request
[10] EAP Request
[10] RADIUS Access Challenge
[11] EAP Response
[11] RADIUS Access Challenge
[12] EAP Success
[12] RADIUS Access Accepted
[13] EAP Key
PMK available
PMK available
Switch to quantum channel
Switch to quantum channel
[14] Send Photons
Switch back to wireless channel
Switch back to wireless channel
SARG04 Quantum Key Distribution
[15] Sifting
[16] Error Correction
[17] Reconciliation
[18] Privacy
Key hierarchy retrieved: PTK, KCK, KEK, TK, GTK
Key hierarchy retrieved: PTK, KCK, KEK, TK, GTK
Start data encryption using TK
Start data encryption using TK

Fig. 1. he QKD based Protocol for Key Exchange

3 Implementation of Agent-Based QKD in WiFi Networks

3.1 Why Multi Agent System?

An Agent can be referred to as a sophisticated computer program, which is capable of acting autonomously to accomplish tasks on behalf of its users, across open and distributed environments. Hence agents have individual internal states and goals, and they act in such a manner as to meet its goals on behalf of its users [20]. Multiple agents can work together to form a multi-agent system (MAS), which offer many advantages over a single agent or centralized approach [16].

Multi Agent Systems in our quantum based key distribution in IEEE 802.11i networks has various advantages. Firstly, the whole protocol can be subdivided into smaller independent modules: 802.11 Association, 802.1X Authentication, Quantum Communication and SARG04 key extraction. These sub-modules can be represented by individual agents to accomplish the main task required. By this way the workload can be distributed among the sub-modules, rather than handling by a single piece of software (centralized approach). Secondly, there are different varieties of EAP types in use for 802.1X authentication such as EAP-TLS, EAP-TTLS, PEAP etc. Therefore, rather than having separate communication flows for each of them, wrapper agents can be used to implement those different EAP varieties. Thirdly, the system maintenance becomes easy as the agents can work independently. Whenever a new change is required to the protocol, it can be done without effecting to the other modules. Fourthly, the system is open to extensions due to modularization via agents. For example, imagine a case where a new EAP type introduced to the protocol suite. In such instances, it can be easily incorporated into the agent society via another agent.

Agents also offer the intelligent behavior to the system. This is a special feature where other wireless protocol implementations are lacking. With this feature, the agents can be taught to detect possible adversary attacks.

3.2 802.1X Protocol Standards and Possible Attacks on 802.11

Many research papers have shown security vulnerabilities of 802.1X standard [14]. As an example, we shall discuss two such attacks.

Session Hijacking: It was shown that session hijacking is possible on 802.1X [14]. This is shown in Figure 2. In these types of attacks, an adversary can spoof communication between a legitimate supplicant and the Authenticator till EAP Success message is received. At this point the adversary sends 802.11 MAC disassociate message using Authenticator's MAC address. This causes the legitimate Supplicant to get disassociated from the Authenticator. However, at this moment the Authenticator is not aware that the legitimate supplicant has kicked out, so it still remains in Authenticated state. The adversary takes this opportunity to hijack the session.

Denial of Service Attacks: Both 802.11 and 802.1X protocols are subject to DoS attacks [14]. These DoS attacks happen in several ways. Adversaries can send fake EAPOL Logoff, EAPOL Start and EAP failure messages towards Authenticator causing the system to fail.

3.3 QKD Based MAS Application

In our approach, we split the main functionalities of each of the major phases to be represented by software agents. As identified before, the authentication and key establishment can be split into following main components:

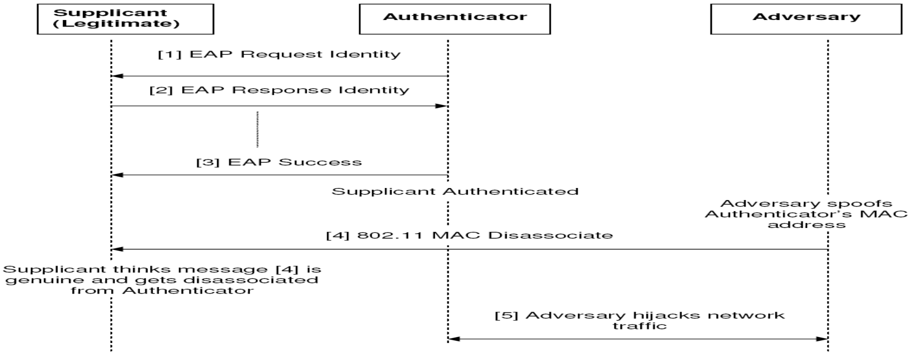

Supplicant (Legitimate)
Authenticator
Adversary
[1] EAP Request Identity
[2] EAP Response Identity
[3] EAP Success
Supplicant Authenticated
Adversary spoofs Authenticator's MAC address
[4] 802.11 MAC Disassociate
Supplicant thinks message [4] is genuine and gets disassociated from Authenticator
[5] Adversary hijacks network traffic

Fig. 2. Session Hijack by MAC address Spoofing

- 802.11 Association and Authentication: for the supplicant to associate with the switch
- 802.1X Authentication: to facilitate mutual authentication between Supplicant and Authenticator
- Quantum transmission: send photons to Authenticator to be used for the key
- Key recovery using SARG04 protocol: recover the final key by removing errors

In our approach, the Agent Society is made up of a main Enterprise as shown in Figure 3 – *The Enterprise*.

Supplicant by executing 802.11 and 802.1X protocols. To facilitate the services, Authenticator spawns a new enterprise for each Supplicant that enters into the wireless network. At the same time the Supplicant too creates one instance of this enterprise to proceed with the communication. Single instance of the enterprise is sufficient at the Supplicant's end as it is only dealing with one Authenticator at a time. These enterprises get together makes the overall Agent Society spanning across the WiFi network served by the Authenticator as shown in figure 3-*The Agent Society*.

802.11 Agent: The main aim of this agent is to perform the 802.11 Association and Authentication. In doing so, this agent can deliver something present 802.11 standard is not capable of doing. That is, with the use of artificial intelligence, this agent is able to take decisions during various adversary attacks.

802.1X Agent: This agent carries out the 802.1X authentication. In this implementation, for simplicity, we only focus on EAP methods that support mutual authentication. This agent is able to support multiple EAP protocols by communicating with different wrapper agents. In addition, it is also able to make decisions on suspicious messages from adversaries similar to what 802.11 Agent does.

Quantum Communication Agent: This agent communicates with hardware devices such as photon transmitter and receiver to make the quantum transmission happen.

SARG04 Agent: The main task of this agent is to execute the SARG04 QKD protocol. It executes the 4 phases of SARG04 protocol in order to extract the final secrete key.

Coordination Agent: The coordination agent communicates with all other agents within the enterprise. Coordination agent in each communication session assures that monitoring efforts and management of internal requests with other agents handled consistently within that specific session.

In this solution, not a single agent is fully aware of the whole communication process. Instead, all agents get together to make the whole communication happen. With this kind of approach, which is quite suitable to be represented as an agent society, modifications can be done effectively.

Similarly, the DoS attack described in above section can be dealt with when 802.1X Agent detects any fake EAP messages.

The software test bed is now being implemented on two computers with one acting as the Supplicant and the other as Authenticator.

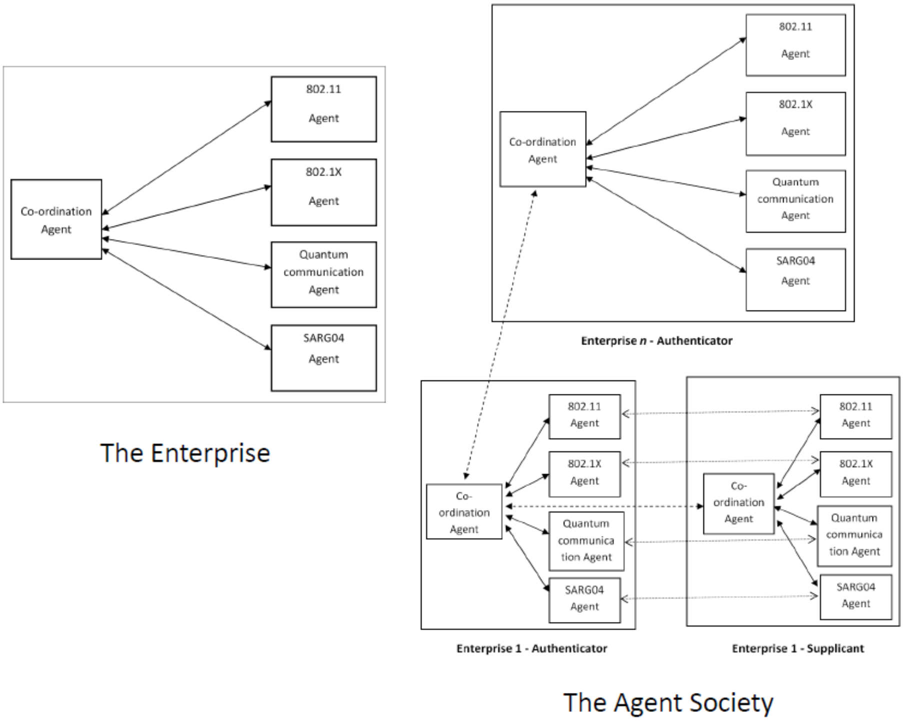


Fig. 3. The Enterprise and The Agent Society

Since the "Native WiFi" software developments are based on C++ platform, we have concentrated on developing MAS application using the same C++ language. We have found that most of the MAS applications only support Java based developments.

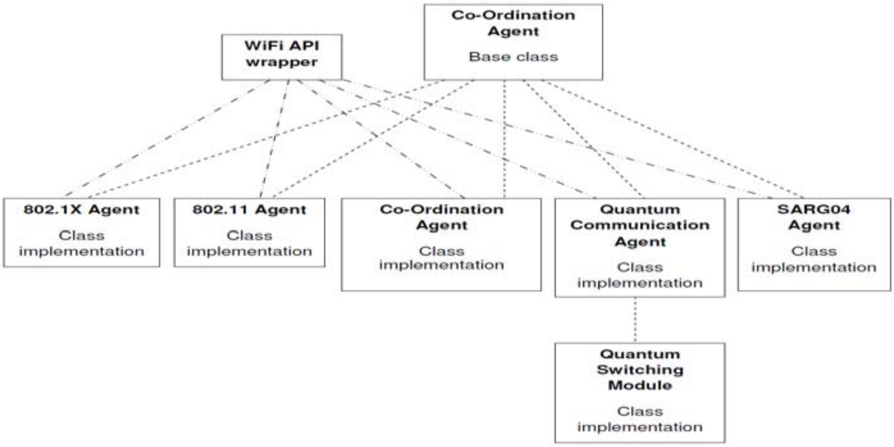


Fig. 4. High level C++ class diagram of the MAS application

Therefore we have decided to write our own application for MAS using C++. This works well with MAS, since C++ being an object oriented language, the agents can easily be represented by C++ classes. As of now we are in the process of developing the SARG04 Agent along with Co-ordination Agent. Appleby and Steward of BT Labs have done a similar approach to prototype a mobile agent based system for controlling telecommunication networks [25].

In this MAS application, we implement C++ class structure to reflect the mobile agents as per the Figure 2. The Co-ordination agent acts as the main class (or agent), which gets created when a Supplicant requires WiFi service. Co-ordination agent calls the other agents only when their service is required. High level C++ class diagram of the MAS application is shown in Figure 4.

4 Conclusion

In this paper we have discussed the use of software agents in QKD based key distribution protocol in WiFi networks. This agent society is particularly useful at Authenticator side as it plays a key role within WiFi networks. As the Authenticator assigns a separate enterprise to look after each Supplicant, the work load can be distributed. This is one of the key requirements for Multi Agent Systems.

This agent approach provides lot of advantages to the wireless communication. Since the key work flows are incorporated into agents, maintenance too becomes easy. Whenever new change to the protocol is needed, it can be done with less effort, without affecting the other agents. It also provides extensibility by allowing different EAP wrapper agents to facilitate different EAP types.

Thus we can conclude that the use of Multi Agent Systems in QKD based WiFi networks offer lot of benefits. There are other research works being done in WiFi area using software agents [21]. We believe our approach using Multi Agent Systems will contribute to develop secure communications for future wireless networks.

References

1. IEEE Std 802.11i, IEEE Standard for Information Technology – Telecommunication and information exchange between systems – Local and metropolitan area networks – Specific requirements. Part 11, Security Enhancements (2004)
2. Huang, X., Wijesekera, S., Sharma, D.: Implementation of QKD in 802.11 Networks. In: Proceeding 2009 IEEE International Conference on Networks Security, Wireless Communications and Trusted Computing, vol. 2, p. 125 (2009)
3. Wijesekera, S., Huang, X., Sharma, D.: Multi-Agent Based Approach for Quantum Key Distribution in WiFi Networks. In: Håkansson, A., et al. (eds.) KES-AMSTA 2009. LNCS (LNAI), vol. 5559, pp. 293–303. Springer, Heidelberg (2009)
4. ANSI/IEEE 802.11, 1999 edn (R2003), Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications (2003)
5. IEEE Std 802.1X, IEEE Standard for Local and metropolitan area networks, Port-Based Network Access Control (2004)
6. Aboba, B., Blunk, L., Vollbrecht, J., Carlson, J., Levkowetz, H.: RFC – 3748, Extensible Authentication Protocol, EAP (2004)

7. Scarani, V., Acin, A., Ribordy, G., Gisin, N.: Quantum cryptography protocols robust against photon number splitting attcks (2004)
8. Bennett, C.H., Brassard, G.: Quantum cryptography: Public-key distribution and coin tossing. In: Proceedings of IEEE International Conference on Computers, Systems and Signal Processing, Bangalore, India, December 1984, pp. 175–179 (1984)
9. Bennett, C.H.: Phys. Rev. Lett. 68, 3121 (1992)
10. Bruß, D.: Optimal Eavesdropping in Quantum Cryptography with Six States. Physical Review Letters 81, 3018 (1998)
11. He, C., Mitchell, J.C.: Analysis of the 802.11i 4-way Handshake
12. De Rango, F., Lentini, D., Marano, S.: Statis and Dynamic 4-Way Handshake Solutions to Avoid Denial of Service Attack in Wi-Fi Protected Access and IEEE 802.11i (June 2006)
13. He, C., Mitchell, J.C.: Security Analysis and Improvements for IEEE 802.11i
14. Mishra, A., Arbaudh, W.A.: An Initial Analysis of the IEEE 802.1X Standard (February 2002)
15. Leavitt, N.: Will IEEE 802.1X Finally Take Off in 2008?, pp. 82–85. IEEE Computer Society, Los Alamitos (2008)
16. Multi-Agent Systems, `http://www.cs.cmu.edu/~softagents/multi.html`
17. SECOQC, Development of a Global Network for Secure Communication based on Quantum Cryptography, `http://www.secoqc.net/`
18. Graham-Rowe, D.: 'Quantum ATM' rules out fraudulent web purchases, New Scientist, Magazine (2629) (November 2007)
19. Mayers, D.: Unconditional Security in Quantum Cryptography. Journal of the ACM 48(3), 351–406 (2001)
20. Software Agents: An Overview, Hyacinth S. Nwana, Intelligent Systems Research, AA&T, BT Laboratories (1996)
21. Automatic Resumption of Streaming Sessions over WiFi Using JADE, Alvaro Suárez, Member, IAENG, M. La-Menza, Elsa M. Macías, Member, IAENG and Vaidy Sunderam
22. Genesereth, M., Fikes, R.: Knowledge interchange format. Version 3.0 Reference Manual, Tech-nical Report Logic 92-1, Computer Science Department, Stanford University (1992)
23. `http://www.idquantique.com/` id Quantique, Quantum Cryptography
24. New Scientist, Quantum ATM rules out fraudulent web purchases, November 10 (2007)
25. Appleby, Steward: Mobile Software Agents for Control in Telecommunications Networks. BT Technological Journal 12(2), 1040113 (1994)

Tolerance Classes in Measuring Image Resemblance*

A.H. Meghdadi[1], J.F. Peters[1], and S. Ramanna[1,2]

[1] Computational Intelligence Laboratory, Dept. Electrical & Computer Engineering,
University of Manitoba, Winnipeg, Manitoba R3T 5V6 Canada
{meghdadi,jfpeters}@ee.umanitoba.ca

[2] Dept. Applied Computer Science, University of Winnipeg, Winnipeg,
Manitoba R3B 2E9
s.ramanna@uwinnipeg.ca

Abstract. The problem considered in this paper is how to measure resemblance between images. One approach to the solution to this problem is to find parts of images that resemble each other with a tolerable level of error. This leads to a consideration of tolerance relations that define coverings of images and measurement of the degree of overlap between tolerances classes in pairs of images. This approach is based on a tolerance class form of near sets that model human perception in a physical continuum. This is a humanistic perception-based near set approach, where tolerances become part of the solution to the image correspondence problem. Near sets are a generalization of rough sets introduced by Zdzisław Pawlak during the early 1980s. The basic idea in devising near set-based measures of resemblance of images that emulate human perception is to allow overlapping classes in image coverings defined with respect to a tolerance ε. The contribution of this article is the introduction of two new tolerance class-based image resemblance measures and a comparison of the new measures with the original Henry-Peters image nearness measure.

Keywords: image resemblance, measure, near sets, perception, tolerance space.

> *An approximation space ... serves as a formal counterpart of pereception ability or observation.*
> – Ewa Orłowska, March, 1982.

* *Intelligent Analysis of Images & Videos* (IAI& V 2009). The insights and suggestions by Homa Fashandi, Christopher Henry, Leszek Puzio, Andrzej Skowron and Piotr Wasilewski concerning topics in this article are gratefully acknowledged. This research has been supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) grants 185986 & 194376, Canadian Arthritis Network grant SRI-BIO-05, and a grant from Manitoba Hydro T277.

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 127–134, 2009.

1 Introduction

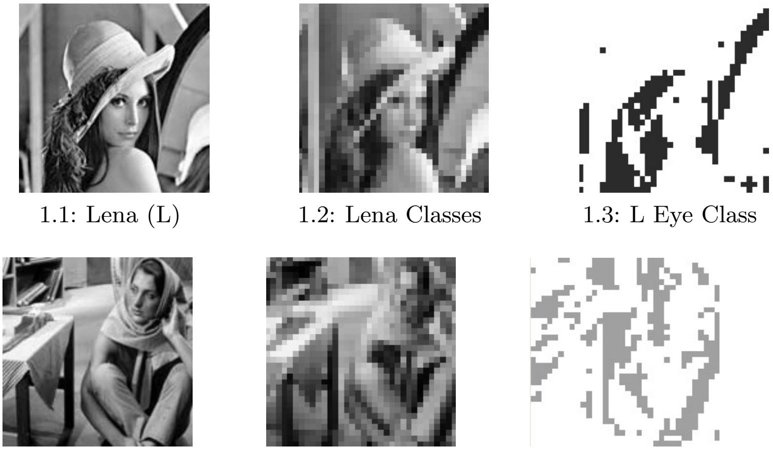
This paper introduces a tolerance near set approach to solving the image correspondence problem, *i.e.*, where one uses image tolerance class matching strategies to establish a correspondence between one or more images. This is one of the central tasks in photogrammetry and computer vision. Recently, it has been shown that near sets can be used in a perception-based approach to discovering correspondences between images (see, *e.g.*, [1,2,3,4]). Sets of perceptual objects where two or more of the objects have matching descriptions are called near sets. Work on a basis for near sets began in 2002, motivated by image analysis and inspired by a study of the perception of the nearness of perceptual objects carried out in cooperation with Z. Pawlak in [5]. This initial work led to the introduction of near sets [6], elaborated in [7,8]. A perception-based approach to discovering resemblances between images leads to a tolerance class form of near sets that models human perception in a physical continuum viewed in the context of image tolerance spaces. A tolerance space-based approach to perceiving image resemblances hearkens back to the observation about perception made by Ewa Orłowska in 1982 [9] (see, also, [10]), *i.e.*, classes defined in an approximation space serve as a formal counterpart of perception.

The term *tolerance space* was coined by E.C. Zeeman in 1961 in modeling visual perception with tolerances [11]. A tolerance space is a set X supplied with a binary relation $\simeq$ (*i.e.*, a subset $\simeq \subset X \times X$) that is reflexive (for all $x \in X$, $x \simeq x$) and symmetric (*i.e.*, for all $x, y \in X$, $x \simeq y$ implies $y \simeq x$) but transitivity of $\simeq$ is not required. For example, it is possible to define a tolerance space relative to subimages of an image. This is made possible by assuming that each image is a set of fixed points. Let O denote a set of perceptual objects (*e.g.*, gray level subimages) and let $gr(x)$ = average gray level of subimage x. Then define the tolerance relation

$$\simeq_{gr} = \{(x, y) \in O \times O \mid \ |gr(x) - gr(y)| \leq \varepsilon\},$$

for some tolerance $\varepsilon \in \Re$ (reals). Then $(O, \simeq_{gr})$ is a sample tolerance space. The tolerance ε is directly related to the exact idea of closeness or resemblance (*i.e.*, being within some tolerance) in comparing objects. The basic idea is to find objects such as images that resemble each other with a tolerable level of error. Sossinsky [12] observes that main idea underlying tolerance theory comes from Henri Poincaré [13]. Physical continua (*e.g.*, measurable magnitudes in the physical world of medical imaging [14]) are contrasted with the mathematical continua (real numbers) where almost solutions are common and a given equation have no exact solutions. An *almost solution* of an equation (or a system of equations) is an object which, when substituted into the equation, transforms it into a numerical 'almost identity', i.e., a relation between numbers which is true only approximately (within a prescribed tolerance) [12]. Equality in the physical world is meaningless, since it can never be verified either in practice or in theory. Hence, the basic idea in a tolerance space view of images, for example, is to replace the indiscerniblity relation in rough sets [15] with a tolerance relation in partitioning images into homologous regions where there is a high likelihood of overlaps, *i.e.*, non-empty intersections between image tolerance classes. The use of image tolerance spaces in this work is directly related to recent work on tolerance spaces (see, *e.g.*, [14,8,16,17,18,19,20,21,22,23]). The contribution of this article is the introduction of two new tolerance space-based image resemblance measures and a comparison of the new measures with the original Henry-Peters nearness measure.

This paper is organized as follows. Sect. 2 presents the basic framework used to define image tolerance spaces. Sect. 3 presents three image resemblance measures. A comparison of the three measures at work is given in Sect. 4.



1.1: Lena (L) 1.2: Lena Classes 1.3: L Eye Class

1.4: Barbara (B) 1.5: Barb Classes 1.6: B Eye Class

Fig. 1. Sample Images and their tolerance classes

2 Tolerance Relations

In general, a *tolerance relation* on a set X in general, is a binary relation that is reflexive and symmetric but not necessarily transitive [12]. A *perceptual tolerance relation* is defined in the context of perceptual systems in (1).

Definition 1. Perceptual Tolerance Relation [17,8]
Let $\langle O, \mathbb{F} \rangle$ be a perceptual system and let $\varepsilon \in \Re$ (set of all real numbers). $\mathbb{F}$ is a set of probe functions $\{\phi_1(x),\ \phi_2(x),\ ...\ \phi_l(x)\}$ where $x \in O$ is a perceptual object and each probe function $\phi_i(x)$ is a real-valued function representing features of physical objects.
For every $\mathcal{B} \subseteq \mathbb{F}$ the perceptual tolerance relation $\simeq_{\mathcal{B},\epsilon}$ is defined as follows:

$$\simeq_{\mathcal{B},\epsilon} = \{(x, y) \in O \times O \mid\ \phi \in \mathcal{B}, \parallel \phi(x) - \phi(y) \parallel \leq \varepsilon\} \tag{1}$$

where $\phi(x) = [\phi_1(x),\ \phi_2(x),\ ...\ \phi_l(x)]^T$ is a feature vector obtained using all the probe functions in $\mathcal{B}$ and $\parallel . \parallel$ is L_2 norm. Corresponding with a tolerance relation $\simeq_{\mathcal{B},\epsilon}$, a tolerance class $x_{/_{\mathcal{B},\varepsilon}}$ related to a perceptual object $x \in O$ is defined as follows:

$$x_{/\sim_{\mathcal{B},\epsilon}} = \{y \in O \mid y \sim_{\mathcal{B},\varepsilon} x\} \tag{2}$$

Example 1. Image Tolerance Classes
Figure 1 shows a pair of images, their tolerance class coverings (Fig. 1.2, Fig. 1.5) and one selected tolerance class relative to a particular image region (Fig. 1.3, Fig. 1.6). Let $\langle O, \mathbb{F}\rangle$ be a perceptual system where O denotes the set of 25×25 subimages. The image is divided into 100 subimages of size 25×25 and can be shown as a set $X = O$ of all the 100 subimages. Let $\mathcal{B} = \{\phi_1(x)\} \subseteq \mathbb{F}$ where $\phi_1(x) = gray(x)$ is normalized average gray scale value of subimage x between 0 and 1. Let $\varepsilon = 0.1$. Observe, for example, the sample tolerance class and containing subimages in Fig. 1.3 corresponding to Lena's left eye. Again, for example, observe the sample tolerance class and containing subimages in Fig. 1.6 corresponding to Barbara's left eye. Relative to the subimage containing Lena's eye and Barbara's eye, each tolerance class contains subimages where the difference between average gray scale values of the subimages and the selected subimage are within the prescribed tolerance level ε. In Sect. 3, separate image tolerance class coverings for each image provide a basis for measuring the degree that pairs of images resemble each other.

3 Tolerance Image Resemblance Measures

This section briefly introduces three tolerance image resemblance measures. In Sect. 4, these two measures will be compared.

3.1 Tolerance Overlap Distribution Measure

A tolerance overlap distribution (TOD) measure is introduced here based on a statistical comparison of the number of overlaps between tolerance classes at each subimage. Suppose $X, Y \in O$ are two images (sets of perceptual objects). The sets of all tolerance classes for image X and Y form a covering for each image.

$$X_{/_{\mathcal{B},\varepsilon}} = \{x_{/_{\mathcal{B},\varepsilon}} \mid x \in X\} \tag{3}$$

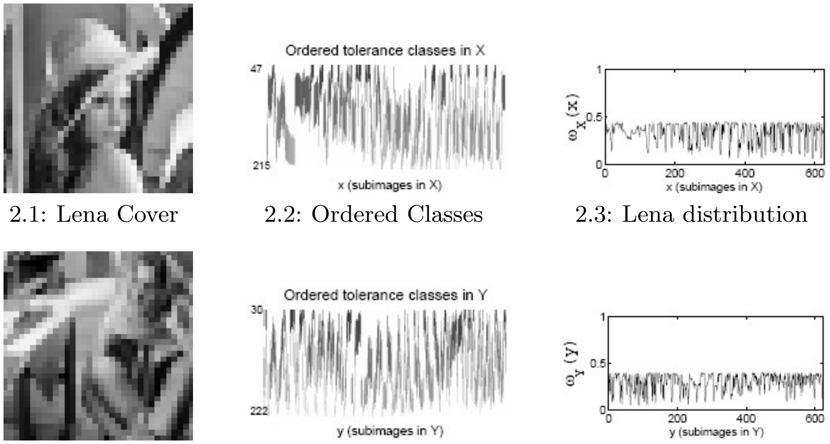
$$Y_{/_{\mathcal{B},\varepsilon}} = \{y_{/_{\mathcal{B},\varepsilon}} \mid y \in Y\} \tag{4}$$

The set of all overlapping tolerance classes corresponding to each object (subimage) x denoted by $\Omega_{X_{/_{\mathcal{B},\varepsilon}}}(x)$ is defined in (5).

$$\Omega_{X_{/_{\mathcal{B},\varepsilon}}}(x) = \{z_{/_{\mathcal{B},\varepsilon}} \in X_{/_{\mathcal{B},\varepsilon}} \mid x \in z_{/_{\mathcal{B},\varepsilon}}\} \tag{5}$$

The degree of overlap at each subimage x can be defined as the normalized number of tolerance classes in $X_{/_{\mathcal{B},\varepsilon}}$ which are overlapping at x denoted by $\omega_{X_{/_{\mathcal{B},\varepsilon}}}(x)$, defined in (6).

$$\omega_{X_{/_{\mathcal{B},\varepsilon}}}(x) = \frac{\left|\Omega_{X_{/_{\mathcal{B},\varepsilon}}}(x)\right|}{\left|X_{/_{\mathcal{B},\varepsilon}}\right|} \tag{6}$$



2.1: Lena Cover 2.2: Ordered Classes 2.3: Lena distribution

2.4: Barb Cover 2.5: Ordered Classes 2.6: Barb distribution

Fig. 2. Tolerance class distributions

Similarly, the degree of overlap between tolerance classes at every subimage $y \in Y$ is denoted by $\omega_{Y_{/\mathcal{B},\varepsilon}}(y)$. Assuming that the set of probe functions $\mathcal{B}$ and the value of ε are known, we use the more simplified notation of $\Omega_X(x)$ and $\omega_X(x)$ for the set $X/_{\mathcal{B},\varepsilon}$ and the notations $\Omega_Y(y)$ and $\omega_Y(y)$ for the set $Y_{/\mathcal{B},\varepsilon}$. Now, let $F_{\omega_X}(\omega)$ and $F_{\omega_Y}(\omega)$ be the empirical cumulative distribution functions (CDF) of the functions $\omega_X(x)$ and $\omega_Y(y)$ respectively, when $x \in X$ and $y \in Y$. To compare the statistical distributions of $\omega_X(x)$ and $\omega_Y(y)$, $|F_{\omega_X}(\omega) - F_{\omega_Y}(\omega)|$ is considered as a measure of difference between distributions at each level of overlap ω. The TOD nearness measure is defined in (7).

$$TOD(X,Y) = 1 - \left(\int_{\omega=0}^{\omega=1} |F_{\omega_X}(\omega) - F_{\omega_Y}(\omega)| d\omega \right)^{\gamma} \tag{7}$$

where $\left(\int_{\omega=0}^{\omega=1} |F_{\omega_X}(\omega) - F_{\omega_Y}(\omega)| d\omega \right)$ represents dissimilarity between distributions. The parameter $\gamma \leq 1$ (or $\gamma \geq 1$) is a scaling factor used when the TOD values are very close to 1 (or 0). A value of $\gamma = 0.6$ is used in this paper.

3.2 Tolerance Nearness Measure

A tolerance nearness measure (tNM) is a variation of the original nearness measure introduced in [2]. The tNM is based on the idea that if one considers the union of two images as the set of perceptual objects, tolerance classes should contain almost equal number of subimages from each image. To see this, assume that X and Y are the sets of perceptual objects (subimages) for a pair of images. Then, $Z = X \cup Y$ is the set of all perceptual objects in the union of images and for each $z \in Z$ defined in (8).

$$z_{/\mathcal{B},\varepsilon} = \{ s \in Z \quad | \quad \|\phi_{\mathcal{B}}(z) - \phi_{\mathcal{B}}(s)\| \leq \varepsilon \} \tag{8}$$

Let $[z_{/\mathcal{B},\varepsilon}]_{\subseteq X}$ denote the tolerance class that is a subset of X and let $[z_{/\mathcal{B},\varepsilon}]_{\subseteq Y}$ denote the tolerance class that is a subset of Y. Then

$$\begin{aligned}
[z_{/\mathcal{B},\varepsilon}]_{\subseteq X} &\triangleq \{x \in z_{/\mathcal{B},\varepsilon} \mid x \in X\} \subseteq z_{/\mathcal{B},\varepsilon} \\
[z_{/\mathcal{B},\varepsilon}]_{\subseteq Y} &\triangleq \{y \in z_{/\mathcal{B},\varepsilon} \mid y \in Y\} \subseteq z_{/\mathcal{B},\varepsilon} \\
z_{/\mathcal{B},\varepsilon} &= [z_{/\mathcal{B},\varepsilon}]_{\subseteq X} \cap [z_{/\mathcal{B},\varepsilon}]_{\subseteq Y}
\end{aligned}$$

Then a tNM is defined in (9) as the weighted average of the closeness between the cardinality (size) of the set $[z_{/\mathcal{B},\varepsilon}]_{\subseteq X}$ and the cardinality of $[z_{/\mathcal{B},\varepsilon}]_{\subseteq Y}$ where the cardinality of $z_{/\mathcal{B},\varepsilon}$ is used as the weighting factor in order to normalize the difference between the size of tolerance classes in each image with respect to the size of the tolerance class that covers both images.

$$tNM(X,Y) = \frac{1}{\sum\limits_{z_{/\mathcal{B},\varepsilon}} |z_{/\mathcal{B},\varepsilon}|} \times \sum_{z_{/\mathcal{B},\varepsilon}} \frac{\min(\ |[z_{/\mathcal{B},\varepsilon}]_{\subseteq X}|\ ,\ |[z_{/\mathcal{B},\varepsilon}]_{\subseteq Y}|\ )}{\max(\ |[z_{/\mathcal{B},\varepsilon}]_{\subseteq X}|\ ,\ |[z_{/\mathcal{B},\varepsilon}]_{\subseteq Y}|\ )} \times |z_{/\mathcal{B},\varepsilon}| \tag{9}$$

3.3 Histogram Image Resemblance Measure

For the sake of comparison, a third image resemblance measure is defined here to compare distributions (histograms) of gray scale values in images without introducing the tolerance spaces. Therefore a histogram similarity measure (HSM) is defined as a measure of similarity between statistical distributions of the "pixels gray scale values" rather than "subimages degree of overlap". Following the same approach given in (7) in the definition of TOD in comparing distributions, the HSM is defined in (10).

$$HSM(X,Y) = 1 - \left(\int_{g=0}^{g=1} |H_{g_x}(g) - H_{g_y}(g)| dg \right)^{\gamma} \tag{10}$$

where g_x and g_y are the normalized gray scales values (g) of pixels in images X and Y, respectively. $H_{g_x}(g)$ and $H_{g_y}(g)$ are cumulative distribution functions of g_x and g_y. γ is the same scaling factor introduced in (7).

4 Tolerance Measures Comparison

This section briefly compares the three tolerance nearness measures introduced in this paper. For conciseness, this comparison is limited to the pair of images given in Fig 1.

Example 2. Sample CDFs for a Pair of Images
Sample image tolerance classes and their overlap distributions are shown in Fig. 3. Vertical axes in figures 2.2 and 2.5 represent tolerance classes in image coverings sorted based on the average gray level among the images in a tolerance class.

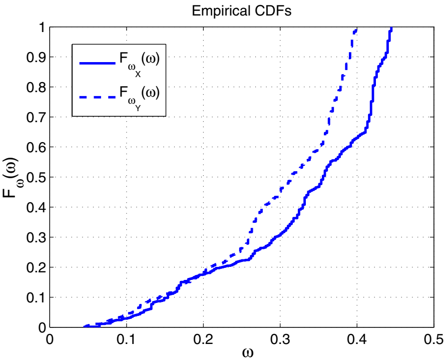


Fig. 3. $\omega_X(x)$ & $\omega_Y(y)$

The empirical CDFs for $\omega_X(x)$ and $\omega_Y(y)$ are shown in Figure 3. The horizontal axis represent the possible values of ω (defined in equation 6) and the vertical axis represent the CDF of the values of ω. The area between the two CDFs is equal to 0.1150, therefore, $TOD = 1-(0.1150)^{0.6} = 0.7268$ (see Fig. 2). Again, for example, let p (subimage size) = 15 and $\varepsilon = 30$. Then, TOD = 0.929, tNM = 0.907 and HSM = 0.875, where the HSM and tNM image resemblance estimates match our intuition about the disparities between the sample images; however, in terms of tolerance class sizes and distributions, TOD and tNM provide a more accurate estimate of image resemblance.

Conclusion

We have presented two tolerance space-based measures of image resemblance. Each of these measures grew out of a study of near sets, a generalization of the Zdisław Pawlak's rough sets. The proposed approach to measuring image resemblance takes its cue from Zeeman's view of tolerance spaces as frameworks for modeling human vision. Basically, we want to measure image resemblance in a manner that, in some sense, mimics our perception. Future work in our study of image resemblance measures will include a consideration of various other distance measures that can be used to gauge the efficacy of the tolerance space-based measures.

References

1. Henry, C., Peters, J.F.: Image pattern recognition using approximation spaces and near sets. In: An, A., Stefanowski, J., Ramanna, S., Butz, C.J., Pedrycz, W., Wang, G. (eds.) RSFDGrC 2007. LNCS (LNAI), vol. 4482, pp. 475–482. Springer, Heidelberg (2007)
2. Henry, C., Peters, J.F.: Near set index in an objective image segmentation evaluation framework. In: GEOgraphic Object Based Image Analysis: Pixels, Objects, Intelligence, University of Calgary, Alberta, pp. 1–6 (2008)
3. Peters, J.F., Wasilewski, P.: Foundations of near sets. Information Sciences 179(18), 3091–3109 (2009)
4. Henry, C., Peters, J.: Perception-based image analysis. Int. J. of Bio-Inspired Computation 2(2) (to appear, 2009)
5. Pawlak, Z., Peters, J.: Jak blisko (how near). Systemy Wspomagania Decyzji I, 57–109 (2002)

6. Peters, J.: Near sets. special theory about nearness of objects. Fundamenta Informaticae 76, 1–27 (2007)
7. Peters, J.F.: Near sets. general theory about nearness of objects. Applied Mathematical Sciences 1(53), 2609–2629 (2007)
8. Peters, J.F.: Tolerance near sets and image correspondence. Int. J. of Bio-Inspired Computation 4(1), 239–245 (2009)
9. Orłowska, E.: Semantics of vague concepts. applications of rough sets. Technical Report 469, Institute for Computer Science, Polish Academy of Sciences (1982)
10. Orłowska, E.: Semantics of vague concepts. In: Dorn, G., Weingartner, P. (eds.) Foundations of Logic and Linguistics. Problems and Solutions, pp. 465–482. Plenum Pres, London (1985)
11. Zeeman, E.C.: The topology of the brain and the visual perception. Prentice Hall, New Jersey (1965); Fort, K.M. (ed.): Topology of 3-manifolds and Selected Topics, pp. 240–256
12. Sossinsky, A.B.: Tolerance space theory and some applications. Acta Applicandae Mathematicae: An International Survey Journal on Applying Mathematics and Mathematical Applications 5(2), 137–167 (1986)
13. Poincaré, H.: The topology of the brain and the visual perception. Prentice Hall, New Jersey (1965); Fort, K.M. (ed.): Topology of 3-manifolds and Selected Topics, pp. 240–256
14. Hassanien, A.E., Abraham, A., Peters, J.F., Schaefer, G., Henry, C.: Rough sets and near sets in medical imaging: A review. IEEE TRansactions on Information Technology in Biomedicine (to appear, 2009)
15. Pawlak, Z.: Rough sets. International Journal of Computer and Information Sciences 11, 341–356 (1982)
16. Peters, J.F.: Discovery of perceptually near information granules. In: Yao, J.T. (ed.) Novel Developements in Granular Computing: Applications of Advanced Human Reasoning and Soft Computation. Information Science Reference, Hersey, N.Y., USA (in press, 2009)
17. Peters, J.F., Ramanna, S.: Affinities between perceptual granules: Foundations and perspectives. In: Bargiela, A., Pedrycz, W. (eds.) Human-Centric Information Processing Through Granular Modelling. SCI, vol. 182, pp. 49–66. Springer, Heidelberg (2009)
18. Bartol, W., Miró, J., Pióro, K., Rosselló, F.: On the coverings by tolerance classes. Inf. Sci. Inf. Comput. Sci. 166(1-4), 193–211 (2004)
19. Gerasin, S.N., Shlyakhov, V.V., Yakovlev, S.V.: Set coverings and tolerance relations. Cybernetics and Sys. Anal. 44(3), 333–340 (2008)
20. Schroeder, M., Wright, M.: Tolerance and weak tolerance relations. Journal of Combinatorial Mathematics and Combinatorial Computing 11, 123–160 (1992)
21. Shreider, Y.A.: Tolerance spaces. Cybernetics and Systems Analysis 6(12), 153–758 (1970)
22. Skowron, A., Stepaniuk, J.: Tolerance Approximation Spaces. Fundamenta Informaticae 27(2/3), 245–253 (1996)
23. Zheng, Z., Hu, H., Shi, Z.: Tolerance Relation Based Granular Space. In: Ślezak, D., Wang, G., Szczuka, M.S., Düntsch, I., Yao, Y. (eds.) RSFDGrC 2005. LNCS (LNAI), vol. 3641, p. 682. Springer, Heidelberg (2005)

Capillary Blood Vessel Tortuosity Measurement Using Graph Analysis

Mariusz Paradowski[1], Halina Kwasnicka[1], and Krzysztof Borysewicz[2]

[1] Institute of Informatics, Wroclaw University of Technology
[2] Department of Rheumatology and Internal Diseases, Wroclaw Medical University

Abstract. Capillaroscopy is a branch of medicine which allows to diagnose various kinds of rheumatic diseases on the basis of observation of visual properties of nail-fold capillaries. Capillaries are tiny blood vessels of various shapes and sizes. Blood vessel tortuosity is one of medical signs. The paper presents a novel blood vessel tortuosity measure designed for capillary analysis. It represents the vessel as a graph and utilizes non-directional and directional traversal algorithms.

1 Introduction

Determination of blood vessel characteristics is an important task in a range of medical diagnostic processes. One of such processes is capillaroscopy diagnosis on the basis of observation of blood vessels in nail-fold skin. These micro blood vessels are called capillaries. There are many blood vessel signs used in capillaroscopy, one of them is blood vessel *tortuosity*. In general, tortuosity is a property of a curve, defining how twisted it is, how many turns it has.

Several ways of tortuosity calculation are presented in literature [1,3,4,5]. They assume that tortuosity is calculated for a single curve (blood vessel). In capillaroscopy we have several degrees of tortuosity [7]. It can manifest itself with single or multiple intersections, or patterns with many different kinds of turnovers located near each other. This means that not only curvature is important, but also a number of intersections. Exemplary capillaries are presented in Fig. 1. A clear and precise vessel segmentation is usually impossible in capillary analysis. Additionally, intersections and branches should be incorporated into tortuosity perception. It may suggest that vessels should be analyzed as a whole, but not as single curves connected to each other. We assume that the proposed method of vessel (nail-fold capillaries) tortuosity measurement should give similar results to capillary classification done by a physician. It should deal with vessels having branches and intersections.

The next section of the paper discusses existing approaches to tortuosity calculation. The third section shows the proposed approach. The fourth section presents the performed experiments. Last section summarizes the paper.

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 135–142, 2009.

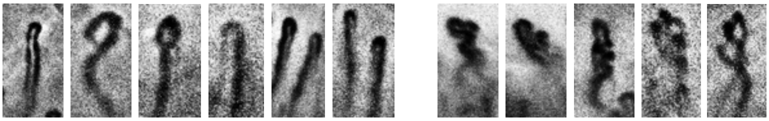


Fig. 1. Exemplary non-tortuous (left) and tortuous (right) capillaries (images acquired at Wroclaw Medical University)

2 Tortuosity Measurement

Various approaches are used for tortuosity measurement: curvature values and changes, arch-chord ratios or vessel angle changes. They require splitting vessel network into curves, and the tortuosity is calculated for each curve.

The tortuosity measurement methods presented in [1] are based on *Arch-chord ratio*, which is defined as a curve length to euclidean distance ratio. However, curves with different visual tortuosity tend to have the same tortuosity coefficient. Grisan et al. [5] proposed *Tortuosity density*, the modification of arch-chord ratio method. Blood vessel (curve) is split into smaller parts, it is done on the basis of points in which a curve starts to bend in different directions. Several approaches are proposed in [3], e.g., *Total curvature*, *Total squared curvature*. They are based on the mathematical definition of curvature, however require blood vessel segmentation into single, non-intersecting, and non-branching parts. Curve angle changes are used in the *Mean tortuosity index* [2]. A set of curve points with an assumed distance between them is selected. Angle changes are calculated on the basis of three successive points. Another method of tortuosity calculation in *3D* is called *Inflection count metric* [4]. It relies on curvature change detection and requires segmentation.

Above tortuosity measures are mainly used to retinal blood vessel analysis, which have different characteristics than capillary blood vessels. Healthy capillaries have a turn, and it does not mean that they are tortuous. It is important to detect how the vessel turns (in which direction) and how it intersects. Grassi [7] points out large influence of branches and intersections on the tortuosity value itself. Automatic segmentation of twisted and intersecting capillary vessels is extremely difficult or even impossible [9,10], therefore all above measures do not fit the needs of capillaroscopy requirements.

3 Proposed Approach - Point-Based Tortuosity

The tortuosity measure method takes as an input a segmented blood vessel, represented as a set of pixels in the image. A single numerical value which represents tortuosity of considered vessel is the output. The proposed approach we call *Point-based tortuosity* τ_p. Non-directed and directed graph analysis, curvature sign calculation and arch-chord ratios are used in the method. Non-directed graph is useful in determination of key points needed to arch-chord ratios calculation. Directed graph analysis allows to order key points and determine the curvature sign. Arch-chord ratios are used to determine local (point) tortuosity. The method goes according to following steps (steps marked with (*) are executed for each pixel in the blood vessel skeleton):

1. Blood vessel skeletonization,
2. Non-directed blood vessel graph construction,
3. Termination and branch points detection,
4. Conversion of the non-directed graph into a directed graph,
5. Key points detection (*),
6. Key points pairing (*),
7. Key point pair tortuosity calculation (*),
8. Best key point pairs determination (*),
9. Best key point curvature sign calculation (*),
10. Tortuosity value averaging.

Above steps are described in the next sections. Steps 1 to 4 are referred as *preparation phase*, steps 5 to 9 – *analysis phase*, step 10 – *synthesis phase.*

3.1 Preparation Phase

In the preparation phase all data necessary for the next phases are calculated using a segmented capillary image. Blood vessel thinning using algorithm presented in [6] allows to extract the skeleton. Next, *termination* and *branch points* are detected. A termination is a point in which the vessel ends. A branch is a point in which it branches or intersects itself. Terminations and branches are detected using 3 × 3 window.

We assume that blood vessel skeleton is a non-directed graph with possible cycles. Each blood vessel skeleton pixel in the image represents a graph node. A graph edge is defined between any two neighboring nodes. Points representing a blood vessel are connected according to *eight-neighborhood* rule. Non-directed graph construction is the basis for directed graph construction, required afterward by the curvature sign determination algorithm. Directed graph has to be constructed in a manner which allows traversal only in one way on the whole vessel. Graph construction starts from an arbitrarily chosen termination point. When a vessel has no termination points (e.g. an ellipse like vessel), one vessel point need to be arbitrarily chosen. A directed graph is generated using *depth-search algorithm* with non-directed graph as an input.

3.2 Analysis Phase

All steps in this phase are repeated for each node S in the vessel skeleton, with node coordinates $S = (x, y)$ as the input.

Step 5 – Key points detection. We search all points of the graph which are placed in the given distance κ from the initial point S. These points, called *key points*, are referred as P. *Breadth-first search* algorithm on the non-directional graph, starting from the given point S, is applied here. It searches for points $P_i \in P$ which meet the following conditions:

- distance from P_i to S has to be larger than κ, but the smallest possible one,
- distance from P_i to S may be less than κ if P_i is a termination point.

The graph traversal goes in different directions. According to *eight-neighborhood* rule there are 8 possible initial traversal directions k, numbered from 1 to 8. Traversal direction are remembered for further processing (*key point pairing*). The initial traversal node gets value 'no direction' ($k = \emptyset$), so $k \in \{\emptyset, 1, 2, 3, 4, 5, 6, 7, 8\}$. Traversal directions are assigned to nodes according to the presented rule, only if they are neighbors to the initial node (x, y). All traversal directions are *inherited* during the graph search.

Graph traversal termination requires graph distance calculation between the current and start nodes. This distance is called *blood vessel distance* (abbreviated as *bvdist*), it is similar to *curve distance* used in other tortuosity measures. Axial neighboring nodes have distance equal to 1. Non-axial neighboring nodes have distance equal to $\sqrt{2}$. In practice such estimation is not enough. To improve this estimation we take into account also *neighbors of neighbors*, the path to a *neighbor of neighbor* node contains one axial and one non-axial edge then the distance is equal to $\sqrt{5}$. The total distance between current and start nodes is the sum of distances along the graph path. In case such approach is insufficient, others estimates may be used, e.g., spline length, after key points are found.

Steps 6, 7, 8. To calculate arch-chord ratio only two nodes (key points) are used, each with different initial traversal direction. Key point selection is based on *arch-chord ratio* [1,5] measurement. Arch-chord ratio is defined as the euclidean distance between nodes to the *blood vessel distance* between these nodes. It is calculated for all key point pairs A, B from P, having different initial traversal directions. Arch-chord ratio value $\tau_s(S, A, B)$, measured using two key points A and B, is defined as follows:

$$\tau_s(S, A, B) = \frac{bvdist(S, A) + bvdist(S, B)}{\sqrt{(A_x - B_x)^2 + (A_y - B_y)^2}}, \tag{1}$$

where: A, B – found key points from P, A_x, B_x, A_y, B_y – x, y coordinates of A and B nodes, $bvdist(A), bvdist(B)$ – blood vessel distance calculated during graph traversal from node S to node A and B, respectively.

From all key points pairs only one (A^*, B^*) is selected, for which arch-chord ratio is minimum:

$$(A^*, B^*) = \arg \min_{A \in P, B \in P} \tau_s(S, A, B), \quad where \quad A_{dir} \neq B_{dir}. \tag{2}$$

Selection of the pair with minimum arch-chord ratio has the following interpretation. Blood vessels are usually smooth and do not bend very rapidly. If two vessels intersects, selecting key points with minimum arch-chord ratio increases the chance of selecting them properly, meaning that both key points belong to the same vessel.

Step 9 – Best key point curvature sign calculation.

Best key point selection results in a pair (A^*, B^*) for which arch-chord ratio is smallest. Those two nodes are selected as local blood vessel approximation. Using those two nodes curvature sign is calculated. Curvature sign plays a crucial role in the presented tortuosity measure. It allows to differentiate between

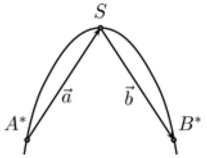


Fig. 2. Data required for curvature sign calculation. Cross product of ***a*** and ***b*** allows sign detection only if key points are ordered in the same way for the whole vessel.

capillaries which turn only one direction and have correct reversed ”U” shape and blood vessels which have many different turns. Curvature sign is calculated using the directed graph, which is determined during the preparation phase. Directed graph allows traversal of the whole vessel in only one direction and this is crucial feature for the calculation. Let us define two *3D* vectors ***a*** and ***b*** using A^* and B^* key points, S coordinates and assume $z = 0$. These vectors are used to determine local curvature sign (see Fig. 2). Using normalized ***a*** and ***b*** vectors, cross product ***N*** is calculated.

$$\boldsymbol{a} = [S_x - A_x^* \; S_y - A_y^* \; 0], \boldsymbol{b} = [B_x^* - S_x \; B_y^* - S_y \; 0], \boldsymbol{N} = \frac{\boldsymbol{a}}{\|\boldsymbol{a}\|} \times \frac{\boldsymbol{b}}{\|\boldsymbol{b}\|}. \tag{3}$$

Sign of z coordinate of vector ***N*** represents curvature sign. Points A^* and B^* have to be ordered in the same manner for the whole vessel, using the following approach. The shortest paths from A^* to S and B^* to S are analyzed. The given below decision rules are applied to calculate the ordering $\eta(A^*, B^*)$, value of 1 represents the correct order, -1 represents the reversed order:

- If path A^* to S contains a child of S, then $\eta(A^*, B^*) = 1$.
- If path B^* to S contains a child of S, then $\eta(A^*, B^*) = -1$.
- If path A^* to S contains a parent of S, then $\eta(A^*, B^*) = -1$.
- If path B^* to S contains a parent of S, then $\eta(A^*, B^*) = 1$.

Parent and child nodes of S are taken from the directed graph. Curvature sign calculation $\sigma(S, A^*, B^*)$ incorporates both vector ***N*** and $\eta(A^*, B^*)$ values:

$$\sigma(S, A^*, B^*) = \begin{cases} 0 & if \quad \boldsymbol{N}_z = 0 \\ 1 & if \quad \eta(A^*, B^*) \cdot \boldsymbol{N}_z > 0 \\ 2 & if \quad \eta(A^*, B^*) \cdot \boldsymbol{N}_z < 0 \end{cases} . \tag{4}$$

3.3 Tortuosity Value Averaging – Synthesis Phase

Tortuosity value averaging is the last step in the proposed method. It aggregates all arch-chord ratios $\tau_s(S, A^*, B^*)$ for all nodes $S \in \Omega$ in the vessel skeleton. In the first step arch-chord ratios are averaged, but only if they have the same curvature sign, i.e., τ_p^1 for $\sigma(S, A^*, B^*) = 1$, and τ_p^2 for $\sigma(S, A^*, B^*) = 2$, according to the formula:

$$\tau_p^i = \frac{1}{|\Omega|} \sum_{S \in \Omega} \tau_s(S, A^*, B^*) | (\sigma(S, A^*, B^*) = i) \wedge (\tau_s(S, A^*, B^*) > \epsilon), \tag{5}$$

for $i = 1$ and $i = 2$ respectively, where τ_s is given by eq. 1, and ϵ is an assumed parameter, it must be a bit greater than 1.

In the second step, we extract information if a blood vessel turns generally only in one direction, or if it turns both. We have τ_p^1 and τ_p^2, the larger one is named τ_h (*higher*), the smaller – τ_l (*lower*). These two tortuosity values are combined into a single *point-based tortuosity* τ_p (eq. 6). It is worth mentioning that τ_h and τ_l also provide important information and can be successfully used as blood vessel features.

$$\tau_h = \max(\tau_p^1, \tau_p^2), \quad \tau_l = \min(\tau_p^1, \tau_p^2), \quad \tau_p = (1 + \tau_h)(1 + \tau_l). \tag{6}$$

4 Experiments

A series of experiments was performed to determine properties of the presented tortuosity measure with two sets of images: synthetic and real clinical (capillary) images. Achieved results are presented and discussed. In all presented experiments $\epsilon = 1.01$ (minimum is 1.0).

4.1 Synthetic Data

Synthesized images represent possible capillary parts, they allow to verify if the visual tortuosity perception is met. The data is processed with $\kappa = 35$, chosen experimentally, as it provides good results for clinical data (shown later).

Fig. 3 presents achieved results. Straight line and + sign have minimum tortuosity ($\tau_p = 1$). L and T are a bit tortuous, due to one sharp turn. An important result from the diagnostic point of view is *snake* tortuosity greater than U, C and O tortuosities. It allows to differentiate between capillaries with one turn (perceived as non-tortuous) and with many different turn.

4.2 Clinical Data

The proposed method should be useful for real clinical data, therefore we have performed experiments using capillary images with carefully hand marked blood vessels. Capillary images, together with diagnoses, come from [7]. Exemplary test images, containing single, marked capillaries, are presented in Fig. 4, they include different capillaries. Calculated tortuosity values are also presented. Proposed tortuosity calculation method was also positively confronted with tortuosity given by a specialized physician on a set of 32 manually segmented, different

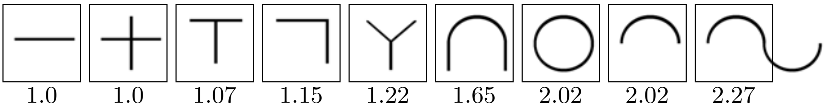


Fig. 3. Synthetic data test. Images shown with point-tortuosity value τ_p.

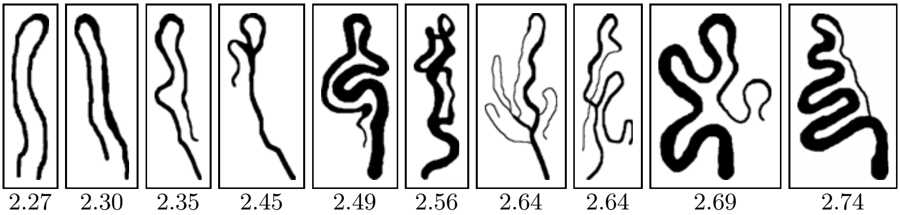


Fig. 4. Exemplary clinical data. Point-based tortuosity τ_p shown for each capillary.

Table 1. Tortuosity classification accuracy with various classifiers and κ values

κ	*NN*	*kNN*	*Bay*	*C4.5*	*MLP*	*LMT*	*avg*	κ	*NN*	*kNN*	*Bay*	*C4.5*	*MLP*	*LMT*	*avg*
5	0.59	0.59	0.72	0.69	0.72	0.81	0.69	10	0.69	0.69	0.81	0.69	0.88	0.88	0.77
15	0.81	0.81	0.81	0.50	0.81	0.75	0.75	20	0.75	0.75	0.84	0.59	0.88	0.81	0.77
25	0.72	0.72	0.84	0.63	0.84	0.88	0.77	30	0.72	0.72	0.84	0.66	0.81	0.84	0.77
35	0.84	0.84	0.88	0.78	0.84	0.88	**0.84**	40	0.75	0.75	0.88	0.81	0.84	0.88	0.82
45	0.72	0.72	0.88	**0.91**	0.78	0.84	0.81	50	0.75	0.75	**0.91**	**0.91**	0.81	0.84	0.83

capillaries. Blood vessels were divided into two classes: *non-tortuous* (17 capillaries) and *tortuous* (15). Naive classification accuracy is equal to 53%. Several classifiers were examined: *Nearest Neighbor* (NN), *k-Nearest Neighbor* (kNN), *Naive Bayes* (Bay), *C4.5 Decision Tree* (C4.5), *Multi-layered Perceptron* (MLP) and *Logit Model Tree* (MLT). *WEKA* [8] package was used as an implementation platform, all classifiers with default parameter values. Taking into account important features of capillaries, we have used following feature vector for classification: total point-based tortuosity τ_p, higher τ_h and lower τ_l directional tortuosity values and the number of blood vessel branches.

Achieved results are presented in Tab. 1. The highest achieved accuracy is 91%. Best averaged quality for $\kappa = 35$ value is 84%. Such results may be perceived as satisfying, especially when comparing to naive classification quality equal to 53%.

5 Conclusions and Future Work

A novel blood vessel tortuosity calculation method is proposed, it is designed for usage with capillaroscopic medical images. The most important features of the method are the following: (i) handling of branching and intersecting blood vessels; (ii) local neighborhood based analysis for better handling of vessels connected with elements of various tortuosity; (iii) usage of curvature sign for better discrimination of blood vessels. Properties of the method are examined in experiments with synthetic images. Clinical data tests are performed on real capillary images. An averaged quality for a set of test images is equal to 84%.

Further research may be focused on better incorporation of branching and intersection information into tortuosity value itself. Branching and intersection are an important aspect in visual tortuosity perception. Blood vessel thickness information may also be considered.

Acknowledgement. This work is partially financed by the Ministry of Science and Higher Education in 2007 - 2009 years as a research project N518 020 32/1454 and a Poland-Singapore joint research project 65/N-SINGAPORE/2007/0.

References

1. Hart, W.E., Goldbaum, M., Kube, P., Nelson, M.R.: Automated measurement of retinal vascular tortuosity. In: Proc. of AMIA Fall Conference, pp. 459–463 (1997)
2. Chandrinos, K.V., Pilu, M., Fisher, R.B., Trahanias, P.E.: Image Processing Techniques for the Quantification of Atherosclerotic Changes. In: Mediterranian Conf. Medical and Bio. Eng. and Computing (1997)
3. Hart, W.E., Goldbaum, M., Kube, P., Nelson, M.R.: Measurement and classification of retinal vascular tortuosity. International Journal of Medical Informatics 53, 239–252 (1999)
4. Bullitt, E., Gerig, G., Pizer, S.M., Lin, W., Aylward, S.R.: Measuring Tortuosity of the Intracerebral Vasculature from MRA Images. IEEE Transactions on Medical Imaging 22(9), 1163–1171 (2003)
5. Grisan, E., Foracchia, M., Ruggeri, A.: A novel method for the automatic evaluation of retinal vessel tortuosity. In: Proc. of 25th IEEE EMBS, pp. 866–869 (2003)
6. Rangayyan, R.M.: Biomedical Image Analysis, Biomedical Engineering Series. CRC Press, Boca Raton (2004)
7. Grassi, W., Del Medico, P.: Atlas of Capillaroscopy, EDRA (2004)
8. Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools and techniques, 2nd edn. Morgan Kaufmann, San Francisco (2005)
9. Kwasnicka, H., Paradowski, M., Borysewicz, K.: Capillaroscopy Image Analysis as an Automatic Image Annotation Problem. In: Proc. of CISIM 2007, pp. 266–271 (2007)
10. Paradowski, M., Kwasnicka, H., Borysewicz, K.: Capillary Blood Vessel Tracking using Polar Coordinates based Model Identification. Advances in Intelligent and Soft Computing, 499–506 (2009)

Image Features Based on Local Hough Transforms

Andrzej Śluzek[1,2]

[1] Nanyang Technological University, School of Computer Engineering,
Blk N4, Nanyang Avenue, Singapore 639798
[2] Nicolaus Copernicus University, Faculty of Physics, Astronomy
and Informatics, ul.Grudziądzka 5, 87-100 Toruń
assluzek@ntu.edu.sg

Abstract. A new method of building local image features is proposed. The features are represented by various shapes (patterns) that can be approximated using Hough transforms. However, the transforms are applied locally (to the current content of a scanning window) so that the shape's location is fixed at the current window's position. Thus, the parameter-space dimensionality can be reduced by two (compared to globally computed Hough transforms) and the transforms can be effectively applied to more complex shapes. More importantly, shapes can be decomposed (two decomposition schemes are proposed) so that the overall complexity of the shapes used as features can be very high. The proposed feature-building scheme is scale-invariant (if scale is a dimension of the parameter space) subject only to diameters of scanning windows.

Keywords: local features, Hough transform, invariant features, shape approximation.

1 Introduction

The Hough transform (its classic form reported in [1]) is a popular tool for a curve fitting over a 2D set of scattered contour points extracted from an image. Although the method is defined for curves of any complexity, it is generally practical only for curves with 2-3 parameters because the number of accumulator bins grows exponentially with the number of parameters. Therefore, in most applications the Hough transform is used to detect straight line segments (that need only two parameters). The modifications proposed for fitting arbitrary shapes also have their limitations since for aspect invariance either additional dimensions (e.g. [2]) or additions of other descriptors (e.g. [3]) are needed.

In this paper, the Hough transform is discussed from a different perspective. Instead applying it globally for the whole image, we use a scanning window (for convenience, circular windows are proposed but windows of any other shape can be alternatively considered) and the curve is fitted to the window's content only, assuming additionally that the curve's location is determined by the window's centre. Therefore, the number of curve parameters can be reduced by 2, and more complex curves can be effectively used.

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 143–150, 2009.

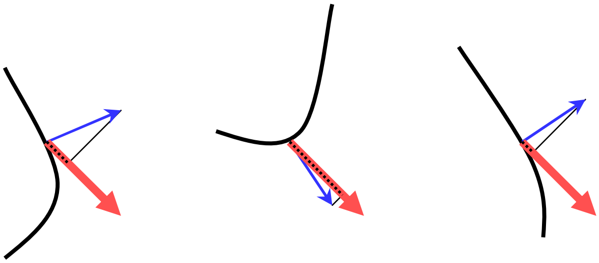


Fig. 1. Three examples of the pixel contributions (the length of dotted lines) to bins of various instances of the curve, depending on the angle between the unit vector normal to the curve (blue arrows) and the gradient vector (red arrows)

An obvious disadvantage of this approach is that any curve can be fitted to almost any window. Therefore, we propose a supplementary mechanism for detecting windows with “the best of best-fitting” curves, i.e. the mechanism for detecting the actual locations of curves within the image.

Another potential disadvantage of the window-based approach is an unpredictable number of contour points within a window (e.g. to few to reliably fit the curve). Therefore, we propose to use all pixels of the window, where each pixel contributes to all accumulator bins. However, the contributed value depends, first, on the pixel’s gradient magnitude and, secondly, on the level of collinearity of the gradient vector with the normal vector of the corresponding instance of the curve (see Fig. 1). The idea is similar to the concept of GWHT (Gradient-Weighted Hough Transform) proposed in [4].

The ultimate objective of using the local Hough transforms is to identify robust local image features, i.e. locations where the selected curves can be fitted to the window’s content with the highest prominence. Such local feature are potentially highly invariant under various photometric and geometric image distortions (including the scale invariance, subject only to the scanning window diameter). Therefore, they can be instrumental in image matching and in retrieval of images with visually similar contents.

In Section 2 we briefly introduce formalisms needed in the proposed method (including explanations on how to localize features based on the local Hough transforms) and explain how the method can be expanded into more complex problems (parallel and sequential decomposition of curves). In Section 3, several typical shapes are discussed and exemplary results are shown.

2 Gradient-Based Local Hough Transforms

In this section, we briefly overview the Hough transforms. The presented model is a modification and improvement of the ideas published previously (e.g. [4] and [5]) so that only the issues that are novel in the proposed method are addressed. Two issues are particularly important: (1) the gradient-based contribution of pixels to the parameter space for curves of any shape and (2) feature localization at local maxima of the Hough transform selectivity.

2.1 Gradient-Based Approach to Hough Transforms

Assume a family of 2D curves described by a parameterized equation

$$f(x, y, p_1, ..., p_n) = 0\,. \tag{1}$$

For practical reasons, we are interested only in curves of semi-C^1 class (i.e. with piecewise-continuous first derivatives).

Consider an image with the intensity function $I(x, y)$ and the gradient vector-function:

$$\nabla_I(x, y) = \begin{bmatrix} \frac{\partial I}{\partial x} & \frac{\partial I}{\partial y} \end{bmatrix}. \tag{2}$$

Any point (x_0, y_0) of the image defines a hyper-surface *HS* in $P_1 \times P_2 \times \ldots \times P_n$ parameter space:

$$HS = \left\{ (p_1, ..., p_n) \middle| f(x_0, y_0, p_1, ..., p_n) = 0 \right\}. \tag{3}$$

However, unlike in standard Hough transforms, we additionally define a weight function *w* over *HS* (i.e. over the family of curves specified by Eq.1) so that for any point $(p_1^A, p_2^A, \ldots, p_n^A) \in HS$:

$$w_{HS}(p_1^A, ..., p_n^A) = \nabla_I(x_0, y_0) \circ norm_{p_1^A, ..., p_n^A}(x_0, y_0)\,. \tag{4}$$

where $\nabla_I(x, y)$ is the image gradient at (x_0, y_0);

$norm_{p_1^A, ..., p_n^A}(x_0, y_0)$ is the unit vector normal to $f(x, y, p_1^A, ..., p_n^A) = 0$ at (x_0, y_0);

and ° represents *dot product* of vectors.

This is the formal model of the effect shown in Fig. 1, i.e. the weight depends on how much the actual gradient at a point is collinear with the normal direction of the curve intersecting the point.

Other steps of using the Hough transform for curve fitting are typical. The parameter space is divided into a suitable number of bins, and each hyper-surface *HS* contributes to all intersected bins the weight w_{HS} computed at the corresponding bin's centre. Subsequently, the winning bins indicate the curves considered the best approximations of the image content.

Although computational complexity of gradient-base Hough transforms is higher, they are superior (as reported in several papers, e.g. [5], [6]) in particular in case of "difficult" images with blurred and/or spur edges, high level of noise, shapes over textured areas, etc.

2.2 Local Hough Transforms

In order to detect local features based on Hough transforms, we have to use the transforms computed locally, e.g. within the content of window scanning the image. The most typical circular windows are used, although any other shape can be considered

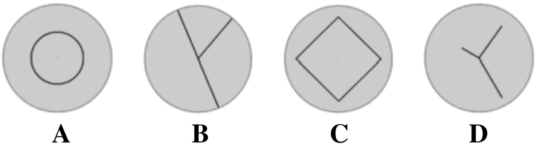


Fig. 2. Predefined location of simple curves within a scanning window; **A** – circles (only one parameter needed, i.e. the radius) needed); **B** – T-junctions (two parameters – orientations of long and short segments); **C** – squares (two parameters – size and orientation); **D** – multi-junctions with unevenly cut arms (two parameters – length and orientation – can be simultaneously applied to all segments in the same 2D parameter space)

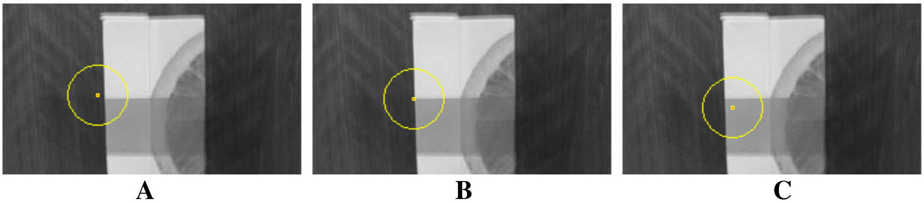


Fig. 3. The scanning window at the actual location of a 90deg T-junction (**B**) and slightly off the actual location (**A** and **C**)

as well. The window cannot be too small in order to provide enough data for a meaningful curve fitting, and larger windows allow detection of features with more diversified sizes so that the scale-invariance (limited only by the window's size) can be achieved. We experimentally established that windows of approx. 25-pixel radius reasonably compromise scale-invariance and computational complexity for a wide range of typical images.

The major advantage of the transforms computed over a scanning window is that we can fix the position of the fitting curve(s) at the current windows location (i.e. the number of curve parameters is reduced by two). Fig. 2 shows how several simple curves can be attached to the window's centre. It can be easily verified that the numbers of parameters for the global Hough transforms of the same curves would be higher (i.e. three for circles, four for T-junctions and squares, and an unspecified number for multi-junctions).

Locally applied Hough transforms cannot localize the best-fit curves within analyzed images (for any location of the scanning window some curve instance can be usually fitted). Thus, we propose a simple yet reliable algorithm based on how selective the locally found fitting curve is.

Assume that the curves of interest are 90deg T-junctions (a special case of a T-junction given in Fig. 2B) and consider three locations of the scanning window within an exemplary image, as shown in Fig. 3. 90deg T-junctions are selected because their local Hough transform needs only 1D parameter space (orientation). Fig. 4 gives the corresponding profiles of the Hough transform in 1D parameter space (the winning orientation values are encircled).

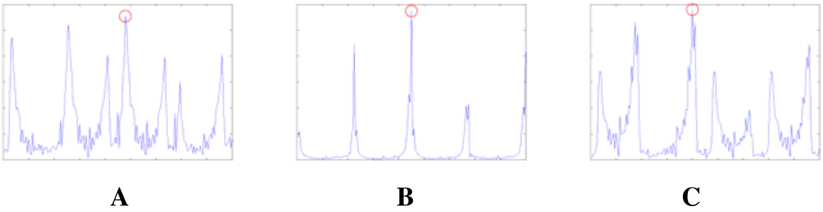


Fig. 4. Profiles of 1D Hough transforms (detection of 90deg T-junctions) for the corresponding window locations given in Fig. 3

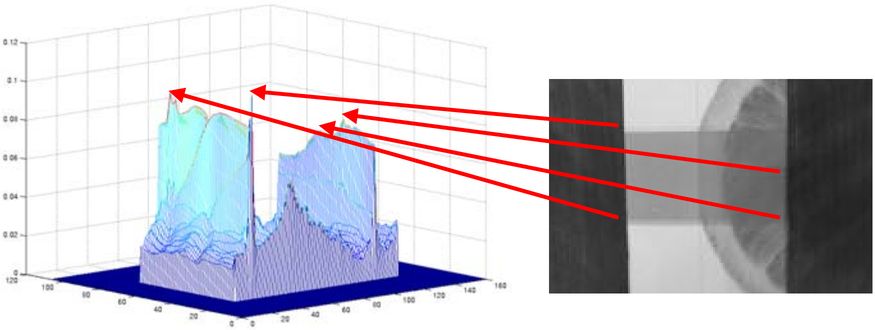


Fig. 5. Correspondences between the local maxima of the selectivity value and the locations of 90deg T-junctions in the analyzed image

The ratio of the winning bin content over the overall content of all bins in the parameter space will be referred to as *selectivity* of the best-fit curve. Formally, *selectivity* of the best-fit curve within the window at (x, y) location is defined as

$$sel(x, y) = \frac{\max\left(v(bin_i)\right)}{\sum_{i=1}^{n} v(bin_i)} \quad i = 1, \ldots, n \,. \tag{5}$$

assuming that the parameter space is divided into n bins ($bin_1, \ldots, bin_n$).

At the locations where the best-fit curves are actually the best approximations of the shapes existing in the analyzed image, the selectivity values reach their local maxima (as illustrated in Fig. 3). The formal proof of this fact exists for curves fully enclosed within the scanning window (the proof is not presented in this paper because of the page number limit). An exemplary image (in which the curve of interest is a 90deg T-junction) and the corresponding selectivity function are shown in Fig. 5. The prominent spikes of the function match the human visual inspection results.

An alternative selectivity measure can be proposed as a ratio between the winning bin content and the integral of the gradient magnitude computed over the scanning window. The results of feature localization using this alternative formula are qualitatively almost the same for all analyzed images, even though they are not identical numerically.

3 Exemplary Local Features

Detection of robust local features is the primary intended usage of the proposed local Hough transforms. Therefore, shapes (curves) typically used as the image features should be of the primary interest. In this section we discuss three important local features: circles, squares and multi junctions (with uneven lengths of their arms) shown in Figs 2A, 2C and 2D, correspondingly.

3.1 Local Hough Transforms for Detecting Circles and Squares

Simple geometric shapes (circles and squares are just typical examples) are very useful features providing important local semantics for image matching and/or image retrieval. Because of simple geometry, the local Hough transform for detection of circles (attached be window's centre) needs only 1D parameters space, while squares need a 2D space (see Figs 2A and 2C). However, multiple copies of the same object can be simultaneously detected in the same parameter space. Exemplary results for circle detection and square detection are shown in Figs 6 and 7.

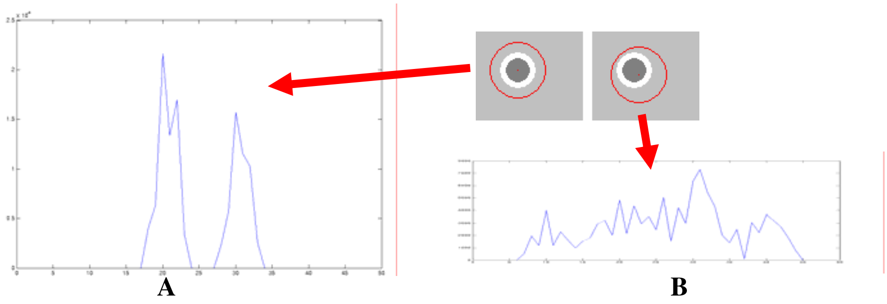


Fig. 6. Profiles of local Hough transforms for circle detection in case of a correctly positioned scanning window (A) and in case of a randomly positioned window (B)

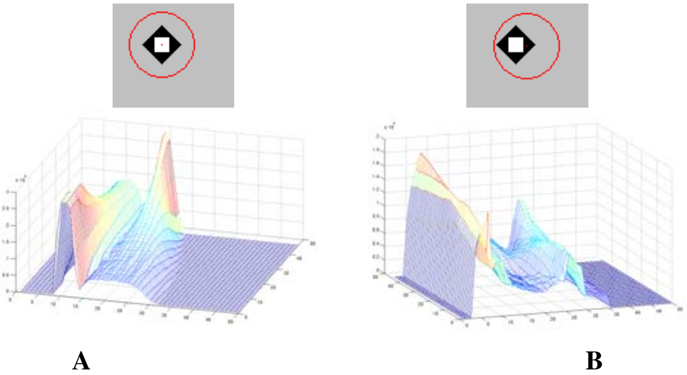


Fig. 7. Profiles of local Hough transforms for square detection in case of a correctly positioned scanning window (A) and in case of a randomly positioned window (B)

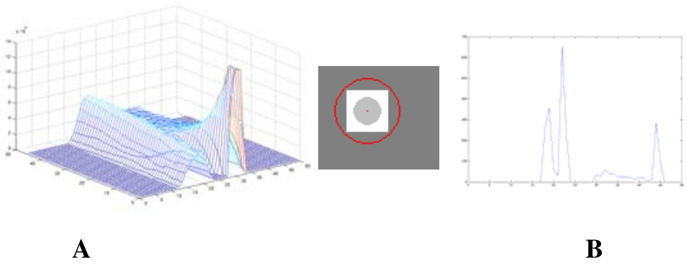


Fig. 8. Profiles of local Hough transforms for a square (A) and circle (B) detection for a window containing both shapes

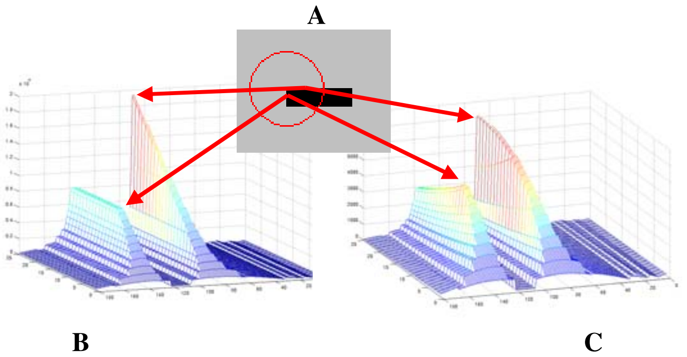


Fig. 9. The local Hough transforms for a two-arm junction (A). The profile is produced without (B) and with (C) the exponential damping function.

In Fig. 6, a window alocated at two concentric circles generates (Fig. 6A) a 2-peak profile in the parameter space, while an incorrectly positioned window produces a number of hardly distinctive spikes (Fig. 6B). Similar effects can be seen in Fig. 7, where the window positioned over two concentric squares (of different orientations, however – Fig. 7A) generates a prominent two-spike profile in the 2D parameter spaces. A randomly positioned window (Fig. 7B) generates a few less distinctive peaks. In both cases (i.e. Fig. 6 and Fig. 7) the selectivity function would have distinctive maxima for the scanning window correctly positioned over the circles/squares. Actually, combinations of squares and circles can be extracted in the same way. Fig. 8 shows results for such a case.

3.2 Detecting Multi-junctions

Multi-junctions (corners are their simplest examples) are probably the most useful local features. Although past papers reported applications of the Hough transform to corner detection (e.g. [7], [8]) the proposed approach is a more general solution.

As shown in Fig. 2D, multi-junctions may have several arms of diversified length so that a 2D parameter space (orientation and length) is needed for the local Hough transform. A simple example of a perfect corner and the profile of its Hough transform are given in Fig. 9.

Fig. 9B shows that the actual length of an arm corresponds to the initial point of a flat segment of the profile. Thus, a simple exponential damping function (decreasing with the length increase) has been incorporated to produce the actual maxima (see Fig. 9C).

4 Final Remarks and Acknowledgements

Because of its limited size, the paper presents only the fundamentals and selected illustrative results of the developed method. Additional details are given in [9] where the method is compared to the alternative moment-base technique. More papers are also currently under review and the publications are expected soon. Other issues of practical significance that should be highlighted are:

- Integration of the method with invariant detectors of interest regions (e.g. DoG, Harris-Laplace or Hessian-Affine);
- Using geometry of extracted local features for geometrically constrained matching of images and their fragments.

Last but not least, the research presented in the paper is a part of A*STAR Science & Engineering Research Council grant 072 134 0052. The financial support of SERC is gratefully acknowledged.

References

1. Duda, R.O., Hart, P.E.: Use of the Hough transformation to detect lines and curves in pictures. Comm. ACM 15, 11–15 (1972)
2. Ballard, D.H.: Generalizing the Hough transform to detect arbitrary shapes. Pattern Rec. 13(2), 111–122 (1981)
3. Artolazabal, J.A.R., Illingworth, J., Aguado, A.S.: LIGHT: Local invariant generalized Hough transform. In: 18th Int. Conf. on Pattern Recognition ICPR 2006, pp. 304–307 (2006)
4. O'Gorman, F., Clowes, M.B.: Finding picture edges through collinearity of feature points. IEEE Trans. Computers 25(4), 449–456 (1976)
5. Van Veen, T.M., Groen, F.C.A.: Discretization errors in the Hough transform. Pattern Recognition 14, 137–145 (1981)
6. Cucchiara, R., Filicori, F.: The vector-gradient Hough transform. IEEE Trans. PAMI 20(7), 746–750 (1998)
7. Davies, E.R.: Application of the generalised Hough transform to corner detection. IEE Proceedings 135(1), 49–54 (1988)
8. Shen, F., Wang, H.: Corner detection based on finding edge points locally. In: Joint Conference on Information Sciences, vol. 6, pp. 773–776 (2002)
9. Sluzek, A.: Building local features from pattern-based approximations of patches: Discussion on moments and Hough transform, EURASIP Journal on Image and Video Processing 2009, ID 959536 (2009)

Capillary Abnormalities Detection Using Vessel Thickness and Curvature Analysis

Mariusz Paradowski[1], Urszula Markowska-Kaczmar[1], Halina Kwasnicka[1],
and Krzysztof Borysewicz[2]

[1] Institute of Informatics, Wroclaw University of Technology
[2] Department of Rheumatology and Internal Diseases, Wroclaw Medical University

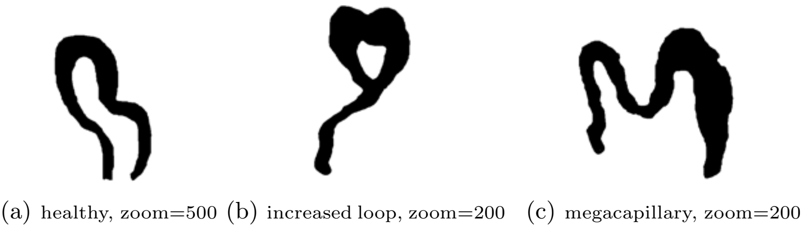
Abstract. The growing importance of nail-fold capillaroscopy imaging as a diagnostic tool in medicine increases the need to automate this process. One of the most important markers in capillaroscopy is capillary thickness. On this basis capillaries may be divided into three separate categories: *healthy*, *capillaries with increased loops* and *megacapillaries*. In the paper we describe the problem of capillary thickness analysis automation. First, data is extracted from a segmented capillary image. Then feature vectors are constructed. They are given as an input for capillary classification method. We applied different classifiers in the experiments. The best achieved accuracy reaches 97%, which can be considered as very high and satisfying.

1 Introduction

Capillaroscopy is a branch of medicine focused on an analysis of nail-fold blood vessels. Abnormalities in blood vessel thickness [1] are one of key medical signs in nail-fold capillaroscopy diagnosis process. An application of computer science methods in nail-fold capillaroscopy is a quite new idea. Some research have been performed around the world, however the results are not satisfactory. Nail-fold capillary measurement research have been performed in the early 2000 and funded by *The Raynauds and Scleroderma Association*. A capillary measurement system (designed for video-capillaroscopy) has been presented in [2]. The system allows to capture video recordings, reduce noises and introduces an amount of automation into the capillary analysis process. User manually selects measurement points and in these points capillary thickness is measured [3].

In the paper we present our approach to the nail-fold capillary classification problem. A single, segmented capillary is an input to the classification task. An output is a class of the input capillary. Classes are defined by medical experts. In order to create input feature vectors for classification, our method combines several pattern recognition techniques, including: capillary skeletonization, curve thickness estimation, local curvature measurement, 2D histograms and object recognition. Feature vectors are extracted on the basis of the capillary thickness and curvature analysis. Because of their importance for successful classification the process of their extraction is described in detail.

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 151–158, 2009.



(a) healthy, zoom=500 (b) increased loop, zoom=200 (c) megacapillary, zoom=200

Fig. 1. Examples of capillaries from the defined capillary classes

The paper is organized as follows. The medical background is presented in the next section. Then we formally present the capillary classification problem. Next, the proposed approach is described in detail. The third section shows the achieved results while the last one summarizes our research.

1.1 Medical Background

One of the most important pathological findings in connective tissue diseases are changes in microcirculation, so called microangiopathia. We have no good clinical, immunological or biochemistry parameters dealing with this problem, so far. Capillaroscopy is a fundamental imaging technique used in the study of microcirculation and seems to be one of the best diagnostic tools for the early detection of microcirculation morphofunctional abnormalities. Microangiopathia is the term strictly connected with enlargement of capillary diameters, forming enlargement loops or megacapillaries.

Enlargement of nail-fold capillaries is the first striking sign of microangiopathy. However, this observation is not true for each case. Microvessels with normal diameter coexist in most instances with definitely enlarged ($> 20\mu m$) or giant loops or megacapillaries ($> 50\mu m$). An increase in capillary diameter can be found in a wide range of conditions, such as systemic sclerosis, dermatomyositis, undifferentiated connective tissue disease, Raynaud‘s phenomen, diabetes mellitus, acrocyanosis. Isolated morphological abnormalities are not unduly rare in the healthy subject. Such changes are homogeneous enlarged loops [4]. Fig. 1 shows exemplary capillaries. Megacapillaries and irregularly enlarged loops are amongst the first morphological abnormalities to be documented in patients with systemic sclerosis. In Raynaud‘s phenomenon patients, single irregularly enlarged loops, even if surrounded by completely normal capillaries, can strongly support the hypothesis of subclinical scleroderma spectrum disorders [5].

If microangiopathy is present, the most likely diagnosis are systemic sclerosis, mixed connective tissue disease, systemic lupus erythematosus and dermatomyositis [4,6]. Capillaroscopy is valuable tool for a correct diagnosis, provided with clinical, serological and immunological findings. It has also prognostic significance in Raynaud‘s phenomenon and scleroderma-pattern disorders [7].

1.2 Problem Definition

The problem of capillary analysis may be formulated as an *image recognition* problem. The analysis should be performed on the basis of a segmented capillary. The result of the analysis is a class of the capillary. There are three possible output classes: *healthy*, *increased loop* and *megacapillary.*

More formally, we assume a set I of pixel coordinates (x and y) representing a single capillary as an input. This means that all points within the set represent one area: $I = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$. The output O is a capillary class: $O = c \in \{\textit{healthy, increased, megacapillary}\}$.

2 Proposed Approach

All steps of the proposed approach are presented in this section. Capillary analysis is the first step, leading to a numerical description of the blood vessel. Blood vessel thickness and curvature are analyzed for the whole capillary skeleton. The second step is data aggregation and feature vector construction. Several approaches are tested, including: thickness averages and standard deviation, thickness and curvature histograms. The last step is the classification phase.

2.1 Capillary Analysis

As we mentioned before, the single segmented capillary is the input. The proposed approach requires a blood vessel skeleton, together with vessel thickness information. Skeletonization is performed according to the algorithm presented in [8].

All points belonging to the extracted skeleton are the basis of further processing. For each point of blood vessel skeleton, thickness and curvature estimation have to be found. Thickness estimation is performed in pixels units. Capillaries are scanned under various zoom settings. Thickness needs to be normalized according to the eq. 1.

$$t = \frac{t_{pixels}}{zoom}, \tag{1}$$

where: t – normalized thickness, t_{pixels} – measured pixel thickness, $zoom$ – scanning zoom factor (values: 50, 100, 200, 500).

Method calibration may be performed using two approaches: setting up device-dependent zoom factors or acquiring the training set on the specific capillaroscopic analysis system.

The measurement of local curvature is performed by *arch-chord ratio* calculation, expressed by eq. 2.

$$c = \frac{d_{euclid}}{d_{path}}, \tag{2}$$

where: d_{euclid} – euclidean distance between two measured points, d_{path} – curve (path) distance between two measured points.

Arch-chord ratio is an efficient way of curvature estimation. It is used, among others in curve tortuosity estimation [9,10]. An exact routine for points selection in arch-chord ration calculation is presented in [10].

A set S of pairs describing a single capillary skeleton is the result of the analysis phase. Each pair refers to a single capillary skeleton point. Formally, the set S can be expressed by eq. 3:

$$S = \{(t_1, c_1), (t_2, c_2), ..., (t_n, c_n)\}, \tag{3}$$

where: n – number of capillary skeleton points, t_k – estimated capillary thickness in skeleton point $k, k \in \{1, ..., n\}$, c_k – estimated capillary curvature in skeleton point $k, k \in \{1, ..., n\}$.

2.2 Data Aggregation

Data generated during the analysis step have to be aggregated in order to create feature vectors, which are the basis of classification. Several approaches to feature vector construction are considered.

Thickness mean and standard deviation. In the simplest approach thickness mean and standard deviation values are applied. All curvature information is discarded in this case. At first glance, it seems to be very simple, however turns to be quite effective.

Fig. 2 shows thickness mean and standard deviation chart. Capillary classes may be separated in an effective way. Healthy capillaries are located in the left bottom part of the chart, capillaries with increased loops are in the middle, megacapillaries are placed on the right side. A single capillary is described by a feature vector (eq. 4), which is composed of two values μ (thickness mean) and σ (standard deviation).

$$f_1 = [\mu, \sigma]. \tag{4}$$

Thickness and curvature histogram. The second approach is based on a thickness and curvature. According to medical information, both of them are important and should be analyzed together.

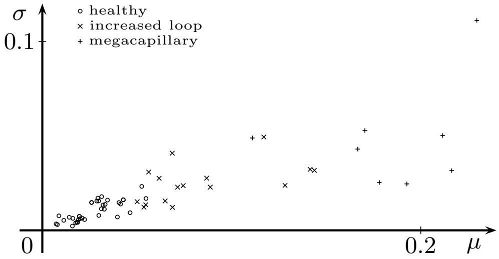


Fig. 2. Thickness mean (μ) and standard deviation (σ) chart, measured for single capillaries

In the first step maximum thickness and curvature values are determined through the dataset. Outliers are detected by checking values against *a priori* given upper threshold. In case values are greater than the threshold, they are set to threshold values (t_{thr} and c_{thr}). It is essential in order to obtain a good values distribution along the histogram. After outliers detection, all values are normalized into $(0, 1)$ range. These preprocessed data are the basis of histogram construction:

$$t_{norm} = \frac{\min(t, t_{thr})}{\max_{t \in S} \min(t, t_{thr})}, c_{norm} = \frac{\min(c, c_{thr})}{\max_{c \in S} \min(c, c_{thr})}. \tag{5}$$

A histogram resolution is another important aspect of the proposed solution. The resolution has to be high enough, because the histogram needs to carry relevant, discriminative information. On the other hand, it has to be low enough, in order to train classifiers correctly.

Examples of thickness and curvature histograms are shown in Fig. 3. Dark areas represent histogram buckets in which most of analyzed point values are located. As it can be noticed, dark values are mostly present in the first column. This means that capillaries are very thin. The first row presents healthy capillaries. The second row shows both increased loop capillaries and megacapillaries. Various configurations of thickness and curvature may be observed.

Another important view of the processed data are per-class histogram averages. Such averages show, how samples are distributed along the whole class. Fig. 4 shows three per-class histogram averages. A clearly dominant feature is capillary thickness, however slightly increased thickness for larger curvature may be observed for *increased loop capillary* class (Fig. 4b). This difference may be an important element for correct discrimination between *increased loop capillaries* (Fig. 4b) and *megacapillaries* (Fig. 4c).

Length of the feature vector depends on the applied histogram resolution and is equal to multiplied thickness and curvature resolutions. The feature vector f_2 is defined as follows:

$$f_2 = [b_{11}, b_{12}, ..., b_{1n}, b_{21}, ..., b_{2n}, ..., b_{mn}], \tag{6}$$

where: b_{xy} represent x, y histogram bin values.

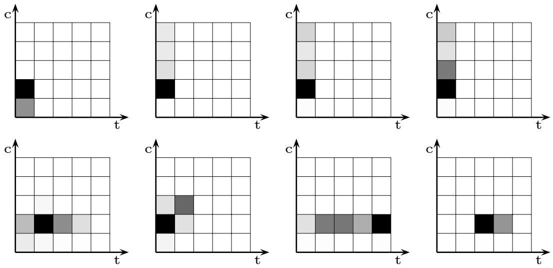
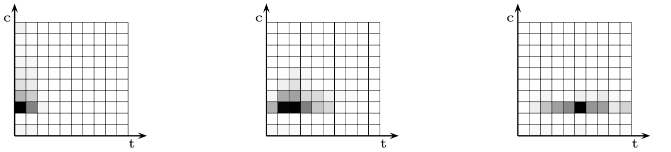


Fig. 3. Thickness (t) and curvature (c) histograms for various capillaries



(a) Healthy (b) Increased loop (c) Megacapillary

Fig. 4. Per-class histogram averages. Histogram resolution is 10 × 10 bins.

Histogram and thickness data (all data). The last approach to features vector creation is a combination of all already presented features. In this case a generated feature vector f_3 contains both *thickness and curvature histogram* information and *thickness mean and standard deviation* values. Its length depends on the histogram resolution (similarly to f_2). The f_3 vector is defined as follows:

$$f_3 = [\mu, \sigma, b_{11}, b_{12}, ..., b_{1n}, b_{21}, ..., b_{2n}, ..., b_{mn}]. \tag{7}$$

Its elements are described by eq. 4 and eq. 6.

2.3 Capillary Classification

As the result of the feature extraction step, each capillary is described by a single, fixed length feature vector for which an appropriate class should be assigned. This means that we deal with a classic classification problem.

The proposed method does not rely on any specific classifier. We applied and evaluated various classification methods during experiments.

3 Experimental Results

The experiments were conducted on a set of segmented, single capillary images. The dataset contains 62 manually segmented capillaries, including: 8 *megacapillaries* and 18 *increased loop* and 36 *healthy* capillaries. Several approaches to classification are tested, including: *Nearest Neighbor*, *k-Nearest Neighbors*, *Naive Bayes*, *C4.5 Decision Tree* and *Logit Model Tree*. It allows to check if proposed feature vectors have general discriminative abilities. Classifiers are taken from *Weka* [11] platform with default parameters.

Three mentioned feature vector construction methods are evaluated. Several histogram resolutions are used (see Tab. 1). To focus on feature vector construction methods, accuracy values for all classifiers have been averaged. In all experiments *leave-one-out* is applied as a cross-validation method.

The achieved results are presented in Tab. 1. Even the simplest thickness based method has reached classification accuracy above 90%. However, histogram based methods show to be even more effective. The most effective approach is a combination of thickness based values and two dimensional histogram. The highest accuracy is equal to 97%. Only several border cases are misclassified. Misclassifications are present between *healthy* and *increased loop* capillaries

Table 1. Classification accuracy versus various 2D histogram resolution settings

Data synthesis	*Res.*	*NN*	*k-NN*	*Bayes*	*C4.5*	*LMT*	*average*
Thickness avg. (f_1)	n.a.	0.92	0.92	0.87	0.90	0.90	0.90
2D histogram (f_2)	5 × 5	0.90	0.90	0.95	0.87	0.92	0.91
2D histogram (f_2)	7 × 7	0.92	0.92	0.90	0.87	0.94	0.91
2D histogram (f_2)	10 × 10	0.92	0.92	0.92	0.85	0.90	0.90
2D histogram (f_2)	15 × 15	0.82	0.82	0.92	0.81	0.85	0.85
All data (f_3)	5 × 5	0.94	0.94	0.95	0.89	**0.97**	**0.94**
All data (f_3)	7 × 7	0.92	0.92	0.89	0.84	0.94	0.90
All data (f_3)	10 × 10	0.95	0.95	0.94	0.94	0.94	**0.94**
All data (f_3)	15 × 15	0.84	0.84	0.92	0.81	0.94	0.87

and between *increased loop* and *megacapillaries*. There are no misclassifications between *healthy capillaries* and *megacapillaries*. These results show the method's ability to cope with the stated problem.

4 Summary

The presented approach focuses on a method of data extraction from a single segmented capillary in order to create feature vector for capillary classification. Three solutions to the feature selection are considered: thickness mean and standard deviation based, joint thickness and curvature histogram and the combination of both of them. In the experiments a number of classifiers are used. Achieved accuracy has reached 97% which may be considered as very high and satisfying.

Further research will concentrate on applying the proposed method into daily practical medical routines. The method will be incorporated into a *Nail-Fold Capillary Analysis System*, which is currently developed by the authors.

Acknowledgements. This work is partially financed by the Ministry of Science and Higher Education in 2007 - 2009 years as a research project N518 020 32/1454.

References

1. Grassi, W., Del Medico, P.: Atlas of Capillaroscopy. EDRA (2004)
2. Allen, P.D., Taylor, C.J., Herrick, A.L., Moore, T.: Image Analysis of Nail Fold Capillary Patterns. In: Proc. of Medical Image Understanding and Analysis, pp. 77–80 (1998)
3. Allen, P.D., Hillier, V.F., Moore, T., Anderson, M.E., Taylor, C.J., Herrick, A.L.: Computer Based System for Acquisition and Analysis of Nailfold Capillary Images. Medical Image Understanding and Analysis (2003)
4. Bollinger, A., Fagrell, B.: Collagen Vascular Disease and Related Disorders. In: Clinical Capillaroscopy, pp. 121–143. Hogrefe and Huber Publishers (1990)
5. Carpentier, P.H., Maricq, H.R.: Microvasculature in systemic sclerosis. Rheum Dis. Clin. North Am. 6, 75–91 (1990)

6. Borysewicz, K., Szechinski, J.: Nailfold Capillaroscopy (NC) evaluation in Systemic Lupus Erythematosus. In: Proc. of 7th European Lupus Meeting, vol. 17(5), p. 88 (2008)
7. Zufferey, P., Deparion, M., Chamot, A.M.: Prognostic significance of nailfold capillary microscopy in patients with Raynaud‘s phenomenon and scleroderma-pattern abnormalities: a six-year follow-up study. Clin. Rheumatol. 11, 536–541 (1992)
8. Rangayyan, R.M.: Biomedical Image Analysis. Biomedical Engineering Series. CRC Press, Boca Raton (2004)
9. Hart, W.E., Goldbaum, M., Kube, P., Nelson, M.R.: Automated measurement of retinal vascular tortuosity. In: Proc. AMIA Fall Conference, pp. 459–463 (1997)
10. Paradowski, M., Kwasnicka, H., Borysewicz, K.: Capillary Blood Vessel Tortuosity Measurement by Directed and Non-Directed Graph Analysis. In: Velásquez, J.D., et al. (eds.) KES 2009, Part II. LNCS (LNAI), vol. 5712, pp. 135–142. Springer, Heidelberg (2009)
11. Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools and techniques, 2nd edn. Morgan Kaufmann, San Francisco (2005)

A Hybrid Method of Biological Computation and Genetic Algorithms for Resolving Process-Focused Scheduling Problems

Ikno Kim and Junzo Watada

Graduate School of Information, Production and Systems
Waseda University
2-7 Hibikino, Wakamatsu-ku, Kitakyushu 808-0135, Japan
octoberkim@akane.waseda.jp, watada@waseda.jp

Abstract. A huge number of different product types are managed through various processes in facilities with different approaches to scheduling. In this paper, we concentrate mainly on process-focused facilities. Sample groups of such facilities and processes were selected: its orders and times were investigated using both biological computation and genetic algorithms. First, biological computation was used to determine practical schedules. Second, genetic algorithms were used to identify which of the schedules determined by biological computation worked best. Here, we examine how combining these methods can be applied to solving process-focused scheduling problems.

1 Introduction

Manufacturing firms and service organisations are continually enhancing their facilities and introducing new machines or tools. When doing so these organisations should update their scheduling operations, or devise new ones, matching their operating processes to customer demands. When dealing with process-focused facilities the corresponding scheduling generates a forward-looking scheme, management of which is complicated. Complexity is a feature of many production facilities [1].

Several researchers have proposed solutions to process-focused scheduling problems that use various meta-heuristic methods, genetic algorithms, and other soft computing methods [2], [3], [4]. However, in this paper, we propose a combination of biological computation and genetic algorithms. This makes it possible to manage a huge number of tasks through the use of molecules, which also make it possibly to handle super-parallel encoding numbers.

To deal efficiently with process-focused facilities in solving scheduling problems, we selected a common case of process-focused schedules. To this we applied, first a biological computation method is created for identifying all feasible schedules, then a genetic algorithms method is created for finding the best schedule from the feasible schedules.

Although the case described in this paper has its own particular features, the model is easily applied to any process-focused schedule because the times and orders are variable. Given this flexibility, the proposed method can be used to measure the efficiency of a large facility using process-focused scheduling.

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 159–165, 2009.

2 Process-Focused Scheduling Problem

A process-focused schedule deals with operating orders and times in process-focused or job shop facilities [1]. The concept of our hybrid method for solving process-focused scheduling problems is shown in Fig. 1.

Process-focused scheduling generates a forward-looking schedule in a high variety, low volume system. It is mostly employed in manufacturing factories and service organisations. The common purpose of these operations is to minimise both average and maximum completion time: the problem is how to achieve this.

The process-focused scheduling problem can be transformed into a directed graph, which is the best way to represent it. The directed graph is composed of pairs and arcs: a pair consists of a job and a machine; an arc connects two different and specific pairs. Fig. 2 shows several cases of process-focused schedule problems represented.

Feasible schedules can be determined by finding a disjunctive arc from each pair, with each of these pairs being acyclic [5]. In the directed graph, if there is a cycle between pairs within a clique it is an impossible schedule, meaning the cycle figures of the conjunctive and disjunctive arcs should not be from different cliques in the directed graph. A subset of the selected disjunctive arcs is existence in the directed graph. Here, this subset can be a feasible schedule in the process-focused facility if, and only if, this directed digraph has no directed cycles. In the directed graph, the pair is denoted as (i, j), where i is the processing step of the machine and j is the processing step of the job.

The maximum completion time is defined as the completion time of the last job that leaves the given system. The longest path from one of the starting pairs to one of the ending pairs corresponds to a feasible schedule. The longest path is composed of a work set, in which the first piece of work starts at 'no time' and the last piece of work finishes at a specified time. Here, each piece of work is followed by either the next on the same machine or the next work of the same job on another machine.

The process-focused scheduling problem is how to minimise the length of the maximum completion time and how to determine a selection of disjunctive arcs. The process here proposed is able to minimise the length of the longest path. Minimising the length of the maximum completion time is obtained by using functions of biological computation that determine all of the acyclic work paths in the directed graph.

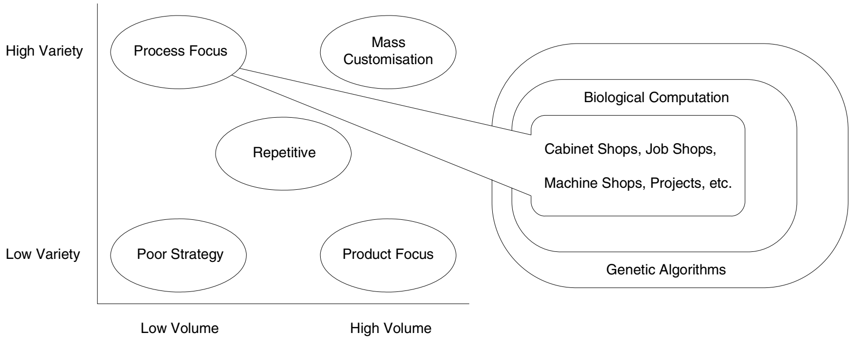


Fig. 1. The concept of the hybrid method for solving process-focused scheduling problems

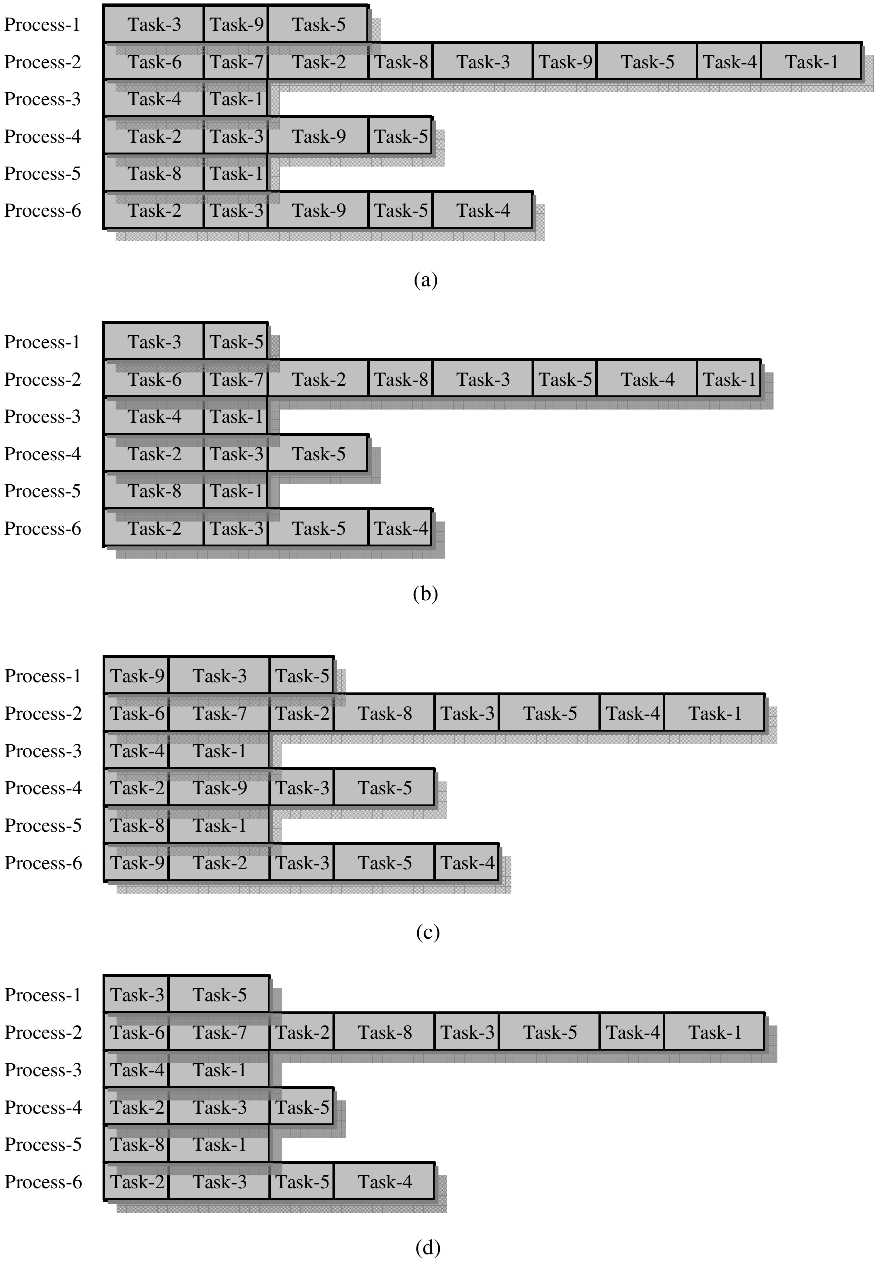

Process-1 Task-3 Task-9 Task-5
Process-2 Task-6 Task-7 Task-2 Task-8 Task-3 Task-9 Task-5 Task-4 Task-1
Process-3 Task-4 Task-1
Process-4 Task-2 Task-3 Task-9 Task-5
Process-5 Task-8 Task-1
Process-6 Task-2 Task-3 Task-9 Task-5 Task-4
(a)
Process-1 Task-3 Task-5
Process-2 Task-6 Task-7 Task-2 Task-8 Task-3 Task-5 Task-4 Task-1
Process-3 Task-4 Task-1
Process-4 Task-2 Task-3 Task-5
Process-5 Task-8 Task-1
Process-6 Task-2 Task-3 Task-5 Task-4
(b)
Process-1 Task-9 Task-3 Task-5
Process-2 Task-6 Task-7 Task-2 Task-8 Task-3 Task-5 Task-4 Task-1
Process-3 Task-4 Task-1
Process-4 Task-2 Task-9 Task-3 Task-5
Process-5 Task-8 Task-1
Process-6 Task-9 Task-2 Task-3 Task-5 Task-4
(c)
Process-1 Task-3 Task-5
Process-2 Task-6 Task-7 Task-2 Task-8 Task-3 Task-5 Task-4 Task-1
Process-3 Task-4 Task-1
Process-4 Task-2 Task-3 Task-5
Process-5 Task-8 Task-1
Process-6 Task-2 Task-3 Task-5 Task-4
(d)

Fig. 2. The four different examples (a) to (d) of process-focused schedule problems. Each of all tasks has its own processing time.

3 Biological Computation Method

A novel method of using biological computation is proposed to identify acyclic work paths, which correspond to feasible schedules within process-focused schedules. In this section, we briefly introduce biological computation, a biological phenomenon, and a molecular encoding process.

3.1 Biological Computation

Biological computation is also called molecular computation, meaning a massive parallel computation. The computation is composed of molecules that constitute deoxyribonucleic acid (DNA), which consist of polymer chains. These are composed of the four nucleotides, adenine, guanine, cytosine, and thymine. This type of computation (sometimes called 'wet computation') is based on the great potential of unique molecular recognitions executed through molecular biological reactions.

When Adleman [6] first proposed molecular computation, he identified a DNA polymerase that has the enzymatic function of copying DNA molecules, which seemed similar to the function of a Turing machine. The DNA polymerase basically composes its complementary DNA molecules using a single strand the helix of a DNA molecule. Its unique characteristic is that if DNA polymerase is properly mixed with a large amount of DNA, the reaction should occur in parallel and at the same time. This suggests the possibility of executing super-parallel processing and creating a huge memory capacity, in contrast to that of conventional electronic computers.

3.2 Watson-Crick Complementarity

DNA consists of strands composed of nucleotides. Each group of all nucleotides is a DNA molecule, and a set of DNA molecules make the structural units of DNA [7].

The four bases of DNA are referred as adenine, guanine, cytosine, and thymine. Adenine bonds only with thymine, while guanine bonds only with cytosine: this phenomenon is called Watson-Crick complementarity.

3.3 Molecular Encoding Process

The example of the directed graph described above, shows all pairs connected at their arcs. The arcs are both conjunctive and disjunctive, but from now on we do not need to distinguish the two, meaning they have the same characteristics.

In the directed graph, the main idea is to encode each DNA sequence based on a binary adjacency matrix, to identify all the acyclic work paths. The directed graph can be easily transformed to a binary adjacency matrix. Each arc of two totally different pairs has its own row and column in the binary adjacency matrix. Assuming n pairs from the directed graph, the rows and columns are labelled $t_{i,j}$, i and $j = 1, 2,\ldots, n$. If there is an arc between two different pairs, then the entry is equal to 1 representing $t_{i,j} = 1$; in contrast, if there is no direction associated with each node, then the entry is equal to 0 representing $t_{i,j} = 0$. For the calculation, two different types of pairs (either 1 or 0), are created to construct an initial library of DNA fragments, and each of the two types has its own row and column label that is defined for molecular encoding pairs in single-stranded DNA.

For the calculation, type 1 denotes an arc that indicates the direction from the pair (w, y) to one of the other connected pairs (x, z). This direction is denoted as a double-encoded substring (w, y)-$3'^{upper}\rightarrow 5'^{upper}$-($x$, z), where both w and x represent the machines, and both y and z represent the jobs.

Second, type 2 denotes complementary sites of DNA molecules between two of the double-encoded substrings. For the directed graph, the forces of hybridisations and ligations make all three different pairs connect to each other. They are thus lined up sequentially together in double-stranded DNA.

4 The Combination Method

The main purpose of this paper is to propose a novel hybrid method, composed of first biological computation and then use of genetic algorithms to resolve process-focused schedule problems. Biological computation [8] is exploited to identify all feasible schedules, and then genetic algorithms are used to find the best of the feasible schedules. In this section, we describe the combined method and simulation-based experiments, and also show the results.

4.1 Biological Computation with Genetic Algorithms

Different types of scheduling decisions are made daily, at every level in different industries and organisations, from manufacturing processes, to project management and task allocation. These decisions commonly appear to be solving large scale combinatorial problems. However, any large scale process has the potential to be extremely difficult to run efficiently: such processes require optimal solution.

The best way to find that solution is to employ two techniques, using biological computation to collect all feasible schedules and then genetic algorithms to discover which schedule is the best.

This combination method provides good and executable solutions, which can be applied to demanding scheduling problems and approach an optimal solution.

4.2 Experiment Procedure

The biological computation proposed is based on a splicing operation. The procedure proposed for its experimental use is as follows:

First, recall the double-encoded substrings for type 1 and complementary sites for type 2, both of which are encoded based on the encoding the directed pairs. In the directed graph, DNA sequences of the existing arcs of the pairs are generated and encoded in single-stranded DNA for hybridisations and ligations.

Second, all of the encoded pairs of the DNA sequences and their complementary sites are artificially synthesized and put into a test tube. They are first heated and then cooled for hybridisation. After hybridisation, all DNA fragment sequences in their own loops are represented as circular DNA sequences, corresponding to cyclical operating schedules. All of these circular sequences should be detected and removed.

Third, the size of the separated DNA strands should be measured using a gel electrophoresis method, which is a biological technique. All the lengths of the hybridised DNA sequences should be measured in this way.

Finally, affinity separation is used with magnetic beads classifying each machine by using the complementary sites of all the pairs. This distinguishes the order of each job in each machine for the process-focused schedules.

4.3 First Step Results

For the first step, biological computation was applied to the splicing operation based on the DNA encoding process for the problem of determining feasible schedules. The pattern of each DNA substring was described by each pair, which corresponded to the DNA sequence. Additionally, the pattern of each complementary site was described by each concatenation of DNA fragments.

Before executing the simulated experiment, all type 1 pairs coming from the given substrings and their complementary sites corresponding to type 2 had to be generated to find the circular DNA fragments. For this encoding process, the fitting restrictive enzymes were added to each of the substrings and complementary sites to separate the circular DNA fragments.

The orders for each job for each machine can be determined by the affinity separation method. The selected job orders for each machine corresponded to all of the possible feasible schedules for the example of several cases. Each selected feasible schedule of all cases had its own maximum completion time. This could be easily measured, when and if the job orders for each machine were exactly determined.

4.4 Second Step Results

In the second step, genetic algorithms were used to identify which of the schedules selected by the first step would be best. For this calculation, the population of the fitness values was set by the change of generations.

The maximum values are the individuals with the highest fitness values in each generation. The average values are all the individuals with average fitness values in each generation. The point here is that the individuals are not distributed in a target optimal area, but are widely scattering in each generation. Table 1 shows the final results of both the maximum completion time and the average completion time for each case.

Table 1. Results of both the maximum completion time and the average completion time for the four example cases

Case 1		Case 2		Case 3		Case 4	
The Average Completion Time	The Maximum Completion Time	The Average Completion Time	The Maximum Completion Time	The Average Completion Time	The Maximum Completion Time	The Average Completion Time	The Maximum Completion Time
184	273	172	226	160	223	151	209

5 Conclusions

In the example, the minimum scheduling time has been obtained by using the combined methods in sequential steps. The first process identified all feasible scheduling times using biological computation, and the second process selected the minimum scheduling time. From the results, we can see that the combination method can be utilised to solve process-focused scheduling problems, as well as helping scheduling decision makers.

For our future study, we plan to develop this method of computation, which can be applied not only for process-focused scheduling problems, but also to other intractable problems.

References

1. Heizer, J., Render, B.: Operations management, 9th edn., pp. 254–279. Pearson Education, Inc., London (2008)
2. Glover, F., Laguna, M.: Tabu search, pp. 25–57. Kluwer Academic Publishers, Dordrecht (1997)
3. Duan, L., Havens, W.S., Dilkina, B.: Applying systematic local search to job shop scheduling problems, pp. 22–31. VDM Verlag Dr. Müller Aktiengesellschaft & Co. KG (2008)
4. Pham, D.T., Karaboga, D.: Intelligent optimisation techniques, pp. 8–15. Springer, Heidelberg (2000)
5. Pinedo, M.: Scheduling, theory, algorithms, and systems, pp. 156–185. Prentice-Hall, Inc., Englewood Cliffs (2002)
6. Adleman, L.: Molecular computation of solutions to combinatorial problems. Science 266, 1021–1024 (1994)
7. Hartl, D., Jones, E.: Essential genetics: A genomics perspective, 3rd edn., pp. 210–242. Jones and Bartlett Publishers, Inc (2005)
8. Watada, J.: DNA computing and its application, computational intelligence: A compendium. In: Fulcher, J., Jain, L.C. (eds.), pp. 1065–1086. Springer, Heidelberg (2008)

Searching Cliques in a Fuzzy Graph Based on an Evolutionary and Biological Method

Ikno Kim and Junzo Watada

Graduate School of Information, Production and Systems
Waseda University
2-7 Hibikino, Wakamatsu-ku, Kitakyushu 808-0135, Japan
octoberkim@akane.waseda.jp,watada@waseda.jp

Abstract. In this paper, a new and systematic approach for the integration of fuzzy-based methods and biological computation, named as an evolutionary and biological method, is proposed for searching cliques in a fuzzy graph. When dealing with a number of nodes in a graph, the most intractable problem is often detecting the maximum clique, which is automatically obtained from finding a solution to the arranged cliques in descending order. The evolutionary and biological method is proposed to identify all the cliques and to arrange them in a fuzzy graph, and then to structure all the nodes in the graph, based on the searched cliques, in different hierarchical levels. This challenging approach, involving the integration of two techniques, provides a new and better method for solving clique problems.

1 Introduction

Advanced information technologies have brought many different examples of difficult problems with complicated data. To solve these kinds of problems, various models and methods have been proposed in different areas. In particular, numerous concepts are often formulated, which become interesting issues on the integrated application side. In accordance with this, we focus on two different methods, fuzzy-based methods and biological computing; and we suggest that these two quite different perspective methods can be integrated to form an adaptable method.

Finding the maximum clique is an intractable problem among several key problems in graph theory [1]. If we obtain all the cliques in descending order using biological computing, the maximum clique is easily known and determined. Moreover, this paper uses a fuzzy graph to structure all the detected cliques, each of which is arranged in similar groups.

Several approaches have proposed applying biological computing to intractable problems. Watada *et al.* [2], [3] proposed a real time optimal scheduling algorithm for solving real time elevator dispatching problems using DNA molecules. Jeng *et al.* [4] proposed the use of DNA melting temperature techniques for cable trench problems.

On the other hand, no one has proposed integrated approaches using a biological computing method. Thus, in this paper, we basically combine fuzzy membership grades dealing with reasoning in relations to employ fuzzy analytical methods and biological computing together.

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 166–173, 2009.

2 Fuzzy Graph

The first process determines specific nodes in each clique in order to accurately understand the fuzzy graph. Therefore, in this section, we describe and analyse a model of a fuzzy graph.

2.1 Model Graph with Fuzziness

As Fig. 1 shows, the fuzzy graph has nodes and edges with fuzzy membership grades. In Fig. 1, a node set is denoted by N, which is represented as the eight nodes, $N = \{1, 2,\ldots, 8\}$. In the fuzzy graph, the connecting lines corresponding to edges with fuzzy membership grades represent the relations.

The terms of arranging nodes with edges are important for determining cliques, including the maximum clique in the fuzzy graph. Hence, all the nodes in each clique need to be determined before structuring those nodes in hierarchical levels.

2.2 Similarity Relations

The fuzzy graph deals mainly with the nodes that imply similarity relations. After determining all the components and cliques in the fuzzy graph, the concept of similarity relations is important to structure the given nodes. These similarity relations should be analysed mathematically [5].

Definition 1. Let ε be a fuzzy relation among nodes on a set N, and the following notions are defined as (1) ε is κ-reflexive if $\forall x \in N$, $\varepsilon(x, x) \geq \kappa$, where $\kappa \in [0, 1]$; (2) ε is irreflexive if $\forall x \in N$, $\varepsilon(x, x) = 0$; and (3) ε is weakly reflexive if for all x, y in N and for all $\kappa \in [0, 1]$, $\varepsilon(x, y) = \kappa \Rightarrow \varepsilon(x, x) \geq \kappa$.

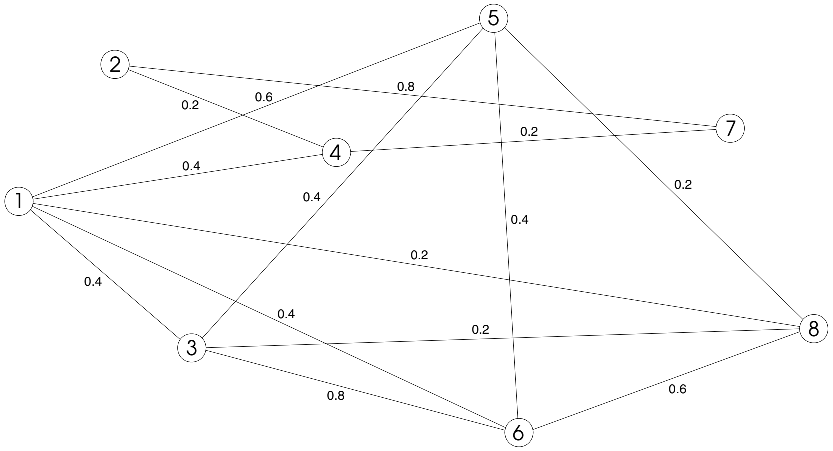


Fig. 1. Example of a fuzzy graph with fuzzy membership grades

Lemma 1. The fuzzy relations among nodes $\varepsilon \circ \varepsilon^{-1}$ should be weakly reflexive and symmetric if ε is a fuzzy relation from N into Q.

Proof. (1) $\varepsilon \circ \varepsilon^{-1}$ is weakly reflexive, which is proved by $(\varepsilon \circ \varepsilon^{-1})(x, x') = \vee\{\varepsilon(x, y) \wedge \varepsilon^{-1}(y, x') \mid y \in Q\} \leq \vee\{\varepsilon(x, y) \wedge \varepsilon(x, y) \mid y \in Q\} = \vee\{\varepsilon(x, y) \wedge \varepsilon^{-1}(y, x) \mid y \in Q\} = (\varepsilon \circ \varepsilon^{-1})(x, x)$; and (2) $\varepsilon \circ \varepsilon^{-1}$ is symmetric, which is proved by $(\varepsilon \circ \varepsilon^{-1})(x, x') = \vee\{\varepsilon(x, y) \wedge \varepsilon^{-1}(y, x') \mid y \in Q\} = \vee\{\varepsilon^{-1}(y, x) \wedge \varepsilon(x', y) \mid y \in Q\} = \vee\{\varepsilon(x', y) \wedge \varepsilon^{-1}(y, x) \mid y \in Q\} = (\varepsilon \circ \varepsilon^{-1})(x', x)$. □

A family of non-fuzzy subsets of nodes is denoted by D^{ε}, which is defined as $D^{\varepsilon} = \{E \subseteq N \mid (\exists 0 < \kappa \leq 1)(\forall x \in N)[x \in E \Leftrightarrow (\forall x' \in E)[\varepsilon(x, x') \geq \kappa]]\}$, where ε is a weakly reflexive and symmetric fuzzy relation among nodes on N. Accordingly, if $D^{\varepsilon}_{\kappa} = \{E \subseteq N \mid (\forall x \in N)[x \in E \Leftrightarrow (\forall x' \in E)[\varepsilon(x, x') \geq \kappa]]\}$ is given, then $\kappa 1 \leq \kappa 2 \Rightarrow D^{\varepsilon}_{\kappa 2} \preccurlyeq D^{\varepsilon}_{\kappa 1}$ can be shown, where '$\preccurlyeq$' is denoted as a covering relation, meaning every node in $D^{\varepsilon}_{\kappa 2}$ is a subset of a node in $D^{\varepsilon}_{\kappa 1}$.

If $\forall x, x' \in N_c$, $\varepsilon(x, x') \geq \kappa$ where N_c is a complete subset of nodes of N, then N_c is called κ-complete with respect to ε. Here, a κ-complete set which is not properly contained in any other κ-complete set of nodes is called a maximal κ-complete set.

Lemma 2. D^{ε} is the family of all maximal κ-complete sets of nodes in a fuzzy graph with respect to ε for $0 \leq \kappa \leq 1$.

Proof. There exists $0 < \kappa \leq 1$ such that $\forall x' \in E$, $\varepsilon(x, x') \geq \kappa$ if $E \in D^{\varepsilon}$ and $x, x'' \in E$, thus $\varepsilon(x, x'') \geq \kappa$ and E is κ-complete. $E \subseteq N_c$ and N_c is κ-complete if N_c is a subset of X where $x \in X$. Let $x \in N_c$, since N_c is κ-complete, $\forall x' \in E$, $\varepsilon(x, x') \geq \kappa$, and since $E \in D^{\varepsilon}$, $x \in E$, thus $N_c \subseteq E$. Hence, E is maximal, then let E be a maximal κ-complete set, and let $x \in X$, then $x \in E \Leftrightarrow \forall x' \in E$, $\varepsilon(x, x') \geq \kappa$ and $E \in D^{\varepsilon}$. □

Lemma 3. There exists some κ-complete set of nodes $E \in D^{\varepsilon}$ such that $\{x, x'\} \subseteq E$, whenever $\varepsilon(x, x') > 0$ is fulfilled.

Proof. If $x = x'$ is fulfilled, $\{x\}$ is certainly κ-complete for $\kappa = \mu_R(x, x)$ where μ_R is a fuzzy set relation R. Let us suppose $x \neq x'$, then $\varepsilon(x, x') = \varepsilon(x', x)$ by symmetry, $\varepsilon(x, x) \geq \varepsilon(x, x')$ and $\varepsilon(x', x') \geq \varepsilon(x, x')$ by weak reflexivity, and also $\{x, x'\}$ is shown as κ-complete where $\kappa = \varepsilon(x, x')$. The family of all κ-complete sets F is denoted by F_{κ}, which has a maximal node set E. Also, this node set is maximal in the family of all κ-complete sets of nodes since any of sets including E, which includes $\{x, x'\}$ as well, thus $E \in D^{\varepsilon}$. □

3 Evolutionary and Biological Method

Detecting specific nodes in each clique is the first step; thus we propose an evolutionary and biological method. In this section, we show this method and introduce a biological phenomenon called Watson-Crick complementarity, which is the main idea for executing the evolutionary and biological method.

3.1 Biological Computation

Biological computation is attracting attention from many scientists, engineers and other researchers [6], [7], [8] because it is a new approach to massively parallel computation and computing basically using bio-molecules, which is a totally different concept from silicon-based computing.

Adleman [9] proposed a computation based on DNA molecules when he found a DNA polymerase that has a specific enzyme function, which can copy and amplify DNA molecules. Biological computation is also called molecular computation because a molecular form of computation can be created based on structure calculations used to make the best of functional, technical, and systematic DNA molecular structures and characters.

The common DNA polymerases compose its complementary sites of DNA molecules. From this unique and regulated function, if a huge number of DNA molecules can be handled in a test tube, the reactions among those DNA molecules must be easily pursued in parallel simultaneously, meaning that the DNA molecules can be expressed as a huge number of data in memory.

3.2 Watson-Crick Complementarity

In DNA, each base is chemically linked to one molecule of the sugar deoxyribose, which is attached to a phosphate group. The computational calculation of DNA is generally executed using these chemical bonding reactions. The four bases of DNA are adenine, guanine, cytosine and thymine. Adenine always bonds with only thymine, while guanine always bonds with only cytosine.

3.3 Clique Detection in Descending Order

The first procedure is to detect all the possible nodes that are implied as cliques. A connectivity matrix can be created from the given nodes, edges and their connections. Fig. 2 shows the connectivity matrix created from the example fuzzy graph.

The connections between the nodes in the model fuzzy graph shown in Fig. 1 can be denoted by $x_{i,j}$, i and $j = 1, 2, \ldots, 8$. The connection values of the strong ties come from node i to node j, and $x_{i,j}$ records which pairs of nodes are adjacent to each other. In the connectivity matrix, if two specific nodes are adjacent, then $0 < x_{i,j} \leq 1.0$; and if two specific nodes are not adjacent, then $x_{i,j} = 0$.

The evolutionary and biological method is proposed for this study based on both the connectivity matrix and the algorithm of the maximal clique problem's solution, which was proposed by Ouyang *et al.* [10]. Our method not only finds all the nodes in cliques, but also all the nodes in components, and arranges the cliques in descending order. The procedure is as follows.

First, nodes in a clique and independent lines are represented by either present or absent in N nodes. If the node is included in the clique or the independent line, then the value is set to 1 in the clique detection, meaning present or adjacent in the graph, otherwise it is 0 in the clique detection, meaning absent.

$$\begin{array}{cc} & \begin{array}{cccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \end{array} \\ \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \end{array} & \left[\begin{array}{cccccccc} 1.0 & 0 & 0.4 & 0.4 & 0.6 & 0.4 & 0 & 0.2 \\ 0 & 1.0 & 0 & 0.2 & 0 & 0 & 0.8 & 0 \\ 0.4 & 0 & 1.0 & 0 & 0.4 & 0.8 & 0 & 0.2 \\ 0.4 & 0.2 & 0 & 1.0 & 0 & 0 & 0.2 & 0 \\ 0.6 & 0 & 0.4 & 0 & 1.0 & 0.4 & 0 & 0.2 \\ 0.4 & 0 & 0.8 & 0 & 0.4 & 1.0 & 0 & 0.6 \\ 0 & 0.8 & 0 & 0.2 & 0 & 0 & 1.0 & 0 \\ 0.2 & 0 & 0.2 & 0 & 0.2 & 0.6 & 0 & 1.0 \end{array} \right] \end{array} \Longrightarrow \begin{array}{cc} & \begin{array}{cccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \end{array} \\ \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \end{array} & \left[\begin{array}{cccccccc} P & A & P & P & P & P & A & P \\ A & P & A & P & A & A & P & A \\ P & A & P & A & P & P & A & P \\ P & P & A & P & A & A & P & A \\ P & A & P & A & P & P & A & P \\ P & A & P & A & P & P & A & P \\ A & P & A & P & A & A & P & A \\ P & A & P & A & P & P & A & P \end{array} \right] \end{array}$$

Fig. 2. Example of a connectivity matrix. 'P' is set to present and included in the clique or the independent line, otherwise 'A' is set to absent and included in the invalid connections.

Second, all the possible combinations between nodes are generated from 2^N using two techniques, which are polymerase chain reaction and parallel overlap assembly.

Third, in the model fuzzy graph, all edges are missing in the original graph, and this graph is called as complementary graph. Any two nodes in the complementary connection are connected in the invalid connection. All of those cliques and independent lines that contain invalid connections are removed by using a restriction enzyme method.

Fourth, the remaining data pool is sorted for selecting the existing DNA-sequences of value 1 after removing all the invalid connections between nodes.

Fifth, all the cliques can be found based on monitoring the lengths of DNA sequences using a gel electrophoresis apparatus. In this step, the shortest DNA strand corresponds to the maximum number of nodes in a clique, meaning the maximum clique, and the second shortest DNA strand corresponds to the second largest maximum number of nodes in a clique.

Sixth, the fourth step is repeated until all the nodes in a clique are obtained. This step is finished when there is only one set of two connected nodes.

Finally, after all the cliques and sets of two nodes are detected, distinguish each of all the nodes which has the same DNA sequences as any other nodes by checking the DNA encodes and becomes a connector between the cliques or sets of two nodes.

3.4 Experimental Method and Results

The evolutionary and biological method uses a simulation-based DNA experiment to basically detect all possible nodes in each clique and component, and arrange them in descending order. This study applied a splicing operation to find all the cliques and components on the basis of our DNA encodes.

Each of all nodes represents the DNA-sequence in a binary number that is either 1 or 0, meaning either present or absent. Two different kinds of DNA sequences are given by a position sequence and a value sequence. The position sequences are used for connecting each node of the DNA-sequence, and the value sequences are used for distinguishing whether or not the position sequences contain a particular node.

The size of the clique is found by monitoring the length of DNA strands based on the simulated DNA experiment. The largest clique of nodes is represented by the shortest length of DNA strand. A set of cliques is denoted by C, and n cliques are labelled C_i, $i = 1, 2,\ldots, n$. The maximum clique of nodes and subset-1 are determined as $C_1 = \{1, 3, 5, 6, 8\}$ and the clique of three nodes and subset-2 are determined as $C_2 = \{2, 4, 7\}$. An independent line of two connected nodes is also determined.

4 Fuzzy-Based Method

The proposed evolutionary and biological method is extended to determine all of the nodes in fuzzy cliques and arrange similarity groups. In this section, we describe how to determine nodes in fuzzy cliques and similarity groups.

4.1 Fuzzy Cliques and Similarity Groups

Let us define a regular graph with a set N_s of nodes and a set E_s of edges to be relation $E_s \subseteq N_s \times N_s$ on a set N_s. A fuzzy relation μ: $N_s \times N_s \rightarrow [0, 1]$ is named a weighted graph or fuzzy graph, and edge $(x, y) \in N_s \times N_s$ has weight $\mu(x, y) \in [0, 1]$. In this case, undirected graphs are considered for simplicity. Thus, the fuzzy relation between nodes in the fuzzy graph is a symmetric relation and all edges in the fuzzy graph are regarded as unordered pairs of nodes [11].

Definition 2. A fuzzy graph is denoted by $G = (\sigma, \mu)$, which is a pair of functions σ: $N_s \rightarrow [0, 1]$ and μ: $N_s \times N_s \rightarrow [0, 1]$, where for all x, y in N_s, then $\mu(x, y) \leq \sigma(x) \wedge \sigma(y)$ is obtained.

Definition 3. A specific fuzzy graph $G_s = (\tau, v)$ is called a fuzzy subgraph of G if $\tau(x) \leq \sigma(x)$ for all $x \in N_s$ and $v(x, y) \leq \sigma(x)$ for all $x, y \in N_s$.

Definition 4. Let G be a fuzzy graph on a set of nodes N_s and the specific fuzzy graph $G_s = (\tau, v)$ be a subgroup, which is induced by $T \subseteq N_s$, where T is a subset of N_s, then G_s is a clique if $(supp(\tau), supp(v))$ is a clique, and G_s is a fuzzy clique if G_s is a clique and every cycle is a fuzzy cycle in (τ, v).

Definition 5. An α-cut of a fuzzy graph that is included in σ and μ for $\alpha \in [0, 1]$, which are the fuzzy sets representing $\sigma_\alpha = \{x \in N_s \mid \sigma(x) \geq \alpha\}$ and $\mu_\alpha = \{(x, y) \in N_s \times N_s \mid \mu_\alpha(x, y) \geq \alpha\}$, where $\mu_\alpha \subseteq \sigma_\alpha \times \sigma_\alpha$ then $(\sigma_\alpha, \mu_\alpha)$ is a graph with the set of nodes σ_α and the set of edges μ_α.

4.2 Detection Results

All the possible nodes in fuzzy cliques are determined, and we define the determined two subsets of all of the two-node combinations as X_{c1} and X_{c2}.

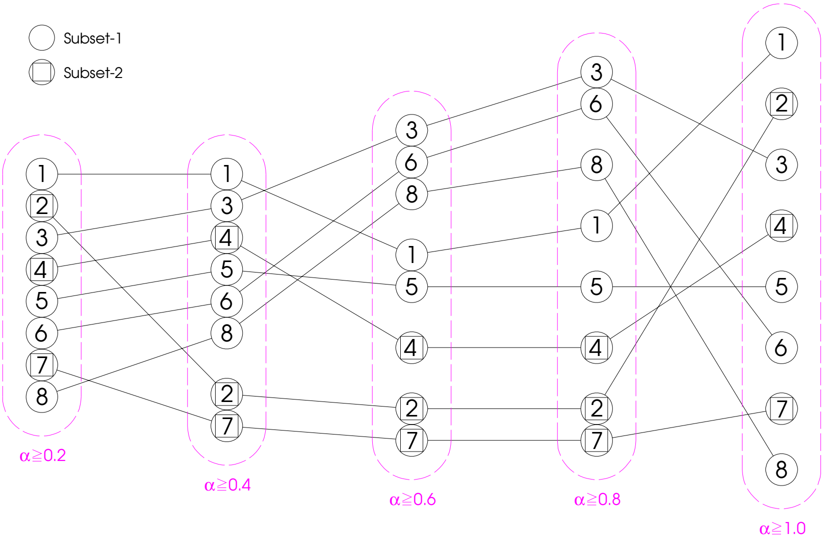


Fig. 3. Similarity hierarchical structure for the two different subsets in the five different levels

In addition, let $\sigma(x) = 1.0$ for all $x_{c1}{\in}N_s$ and $x_{c2}{\in}N_s$ be the fuzzy subset of X_{c1} and X_{c2}, also let each μ be the fuzzy subset of each X_{c1} and X_{c2}. Fig. 3 shows the similarity hierarchical structure where two different subsets of the possible nodes with edges have been arranged into five different levels representing five different α-cuts.

5 Conclusions

Clique groups of nodes have been arranged in descending order based on the two detected subgroups of nodes. Furthermore, in the model fuzzy graph, specific nodes are arranged in fuzzy cliques, which are arranged in similarity groups. In addition, biological computing and fuzzy-based methods have been combined to show how to implement real fuzzy membership grades between nodes with biological computing.

The new integrated approach combines the proposed evolutionary and biological method with the fuzzy-based methods shown in this paper. Moreover, we selected a simple example of a fuzzy graph, which implicitly can be adapted to a large volume of nodes using the characteristics of parallel computation with DNA molecules.

References

1. Bomze, I.M., Pelillo, M., Stix, V.: Approximating the maximum weight clique using replicator dynamics. IEEE Transactions on Neural Networks 11, 1228–1241 (2000)
2. Watada, J., Jeng, D.J.-F., Kim, I.: Application of DNA computing to group control of elevators. In: Iasi, A.I. (ed.) 2005 Anniversary Symposium on the Romanian Society for Fuzzy Systems, A.I., Iasi, Romania, pp. 9–19 (2005)

3. Watada, J.: DNA computing and its application. In: Proceedings of the 8th International Conference on Intelligent Systems Design and Applications, Keynote Speech, Kaohsiung, Taiwan (2008)
4. Jeng, D.J.-F., Kim, I., Watada, J.: Bio-inspired evolutionary method for cable trench problem. International Journal of Innovative Computing, Information and Control 3(1), 111–118 (2006)
5. Rosenfeld, A.: Fuzzy graphs, fuzzy sets and their applications. In: Zadeh, L.A., Fu, K.S., Shimura, M. (eds.), pp. 77–95. Academic Press, New York (1975)
6. van Noort, D., Landweber, L.F.: Towards a re-programmable DNA computer. In: Chen, J., Reif, J.H. (eds.) DNA 2003. LNCS, vol. 2943, pp. 190–196. Springer, Heidelberg (2004)
7. Rose, J.A., Hagiya, M., Deaton, R.J., Suyama, A.: A DNA-based in vitro genetic program. Journal of Biological Physics 28, 493–498 (2002)
8. Sakamoto, K., Kiga, D., Komiya, K., Gouzu, H., Yokoyama, S., Ikeda, S., Sugiyama, H., Hagiya, M.: State transitions by molecules. Biosystems 52, 81–91 (1999)
9. Adleman, L.: Molecular computation of solutions to combinatorial problems. Science 266(11), 1021–1024 (1994)
10. Ouyang, Q., Kaplan, P.D., Liu, S., Libacher, A.: DNA solution of the maximal clique problem. Science 278, 446–449 (1997)
11. Nair, P.S., Cheng, S.-C.: Cliques and fuzzy cliques in fuzzy graphs. In: IFSA World Congress and 20th NAFIPS International Conference, pp. 2277–2280 (2001)

A Biologically Intelligent Encoding Approach to a Hierarchical Classification of Relational Elements in a Digraph

Ikno Kim and Junzo Watada

Graduate School of Information, Production and Systems
Waseda University
2-7 Hibikino, Wakamatsu-ku, Kitakyushu 808-0135, Japan
octoberkim@akane.waseda.jp, watada@waseda.jp

Abstract. Parallel processing functions using molecules have advantages to be exploited for classifying the given relational elements in a digraph. For instance, hierarchical structural modelling is used for classifying complicated objects into a hierarchical structure. In this paper, we consider the example of a digraph of hierarchical structural modelling that can be transformed to sequences of molecules, and propose a biologically intelligent method of encoding molecular sequences of different types, through the hierarchical classification of hierarchical structural modelling. Moreover, we show that this innovative biologically intelligent encoding method can be applied, not only to hierarchical structural modelling, but also to other relational problems composed of elements from digraphs.

1 Introduction

A method of structural modelling of problematic situations is a classification support tool for structuring complicated systems [1], [2]. Structural modelling of problematic situations can be called hierarchical structural modelling for hierarchical classification of relational elements in this study. Hierarchical structural modelling is often used for classifying complicated objects into groups to construct a hierarchically restructured digraph, which provides comprehensible results. Complicated objects may be represented as a set of relational elements in a digraph. It is possible for hierarchical structural modelling to be a key tool of the problem-solving process when the same problems have conceptualising interactions in different organisations [3].

The remaining problem in the electronic calculation of hierarchical structural modelling is associated with dealing with a large number of relational elements. Constructing a hierarchically restructured digraph with many relational elements is problematic because both the calculation time and the number of relational elements increase at the same time.

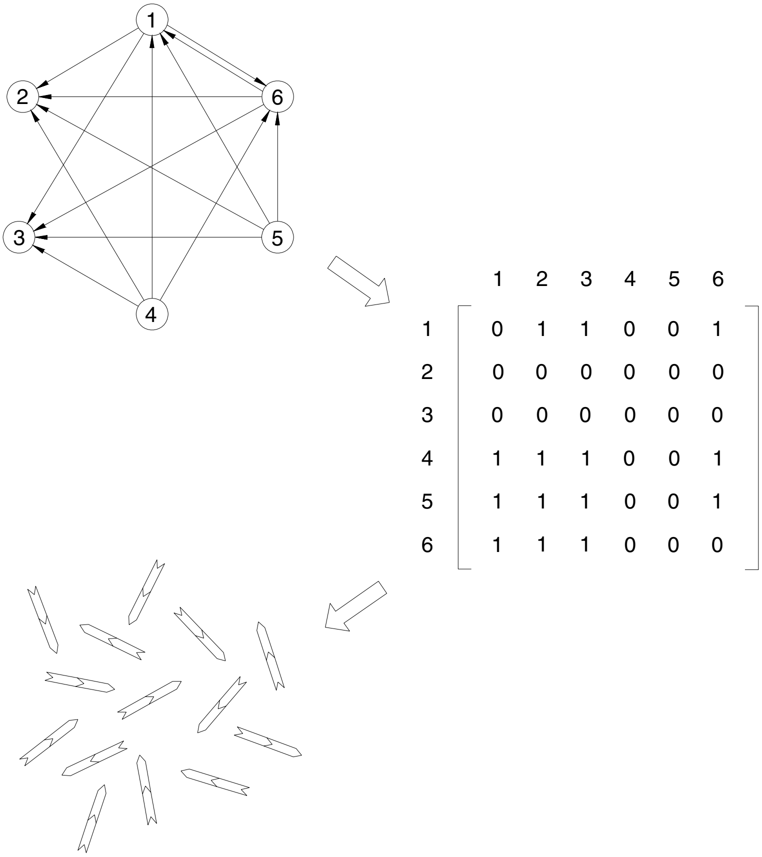
Many researchers [4], [5], [6], [7], [8] have drawn attention to molecular computation, since Adleman [9] first proposed the possibility. In this paper, we show how the capabilities of molecular computation can be expanded. A biologically intelligent method of encoding DNA sequences composed of different types is proposed for a

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 174–180, 2009.

hierarchical classification of relational elements in a digraph, via an example of hierarchical structural modelling. Further, we show how the encoding method can be applied to other relational problems composed of elements and represented as a digraph. This new biologically intelligent encoding method therefore has immense potential.

2 Method of Hierarchical Classifications

One way of classifying complicated objects in a hierarchical classification is through hierarchical structural modelling, in which the given complicated objects become relational elements. The main process of hierarchical structural modelling is to create a digraph composed of relational elements, meaning the sets of relational elements are exploited to construct a digraph.



$$\begin{array}{c|cccccc|}
 & 1 & 2 & 3 & 4 & 5 & 6 \\
\hline
1 & 0 & 1 & 1 & 0 & 0 & 1 \\
2 & 0 & 0 & 0 & 0 & 0 & 0 \\
3 & 0 & 0 & 0 & 0 & 0 & 0 \\
4 & 1 & 1 & 1 & 0 & 0 & 1 \\
5 & 1 & 1 & 1 & 0 & 0 & 1 \\
6 & 1 & 1 & 1 & 0 & 0 & 0
\end{array}$$

Fig. 1. The example representation of a digraph, a binary adjacency matrix, and encoded DNA fragments

To represent a digraph, let us assume n given relational elements are denoted by r_1, r_2,…, r_n, and a set of these relational elements is denoted by R. The relational elements correspond to the relational vertices on the digraph. Fig. 1 shows (1) an example digraph of hierarchical structural modelling; (2) a binary adjacency matrix of hierarchical structural modelling; and (3) encoded DNA fragments. These have been created to describe the biologically intelligent encoding method. In the example digraph, the six specific vertices are represented by six relational elements that are labelled r_i, $i = 1, 2,…, 6$, and a set of the six relational element vertices representing $R = \{r_1, r_2,…, r_6\}$, with considerable arcs between the given relational elements.

3 Molecular Computation Method

An encoding method is basically executed through DNA oligonucleotides on the basis of molecular computation. Molecular computation is strongly associated with DNA oligonucleotides. A key point for molecular computation, also called a biological phenomenon or Watson-Crick complementarity, is to execute the main part of the encoding method in DNA oligonucleotides [10].

An integrated circuit is composed of semiconductors, which are used in electronic computations, and is based on silicon microchip technologies. On the other hand, DNA molecules are referred to in terms of information storage media, and the functions of biological and chemical reactions, caused by various enzymes and proteins. The best way of using our understanding of DNA molecular structures, characters, and functions is to create a computational molecular form on the basis of those structures and architectures. This type of computation is called molecular computation, or biological computation.

4 Biologically Intelligent Encoding Method

The biologically intelligent method basically implements DNA sequences. These can be its specific DNA sequences for each relational vertex in a digraph of hierarchical structural modelling. Thus, the relational vertices from the example digraph can also be transformed into its own DNA sequences. However, the best way of expressing each of the relational vertices is to use a binary adjacency matrix. Such a matrix is easily developed from the digraph. In this section, we introduce the binary adjacency matrix, and describe our biologically intelligent encoding method by showing the different types of DNA sequence.

4.1 Binary Adjacency Matrix

In the biologically intelligent encoding process, the binary adjacency matrix is not needed to raise the resulting matrix to successive powers based on Boolean algebra. The binary adjacency matrix is directly used for the encoding process.

For the model digraph, as shown in Fig. 1, the binary adjacency matrix of size $n \times n$ is six rows and six columns, and has a set of relational element vertices $R = \{r_1, r_2,…, r_6\}$. Each directional order of two relational element vertices has its own row and column, and the rows and columns are labelled $z_{i,j}$, i and $j = 1, 2,…, 6$ in the binary adjacency matrix. If there is an arc from a relational element vertex r_i to a relational element vertex r_j, then $z_{i,j} = 1$, and $z_{i,j} = 0$.

4.2 Encoding Types for Chain and Cyclic Connections

For the biologically intelligent encoding method, eight different encoding types are created to construct an initial library of DNA fragments. As shown in Fig. 2, type C is encoded to identify all of the both chain and cyclic connections. Type C is composed of types CH and CY, where type CH corresponds to chain connections and type CY corresponds to cyclic connections. Type C is a double-encoded substring that represents two different single relational element vertices where an arrowed line indicates the direction from the relational element vertex r_i to the relational element vertex r_j. The digraph includes either a single direction or a set of two parallel arrowed lines. Here, the set of two parallel arrowed lines indicates opposite directions between the relational element vertex r_i and the relational element vertex r_j.

Type C is created to determining types CH and CY. In the binary adjacency matrix, all the row and column labels are denoted by i and j, and the entries of the row and column labels are defined as

$$z_{i,\,j} = 1 \text{ for } i \text{ and } j = 1, 2, \cdots, n, i \neq j, \text{ and all}(i,\ j) \in Z\ . \tag{1}$$

After the hybridisation and ligation process of all the double-encoded substrings and their complementary sites, if there any circular DNA fragments are detected, then the circular DNA fragments correspond to either type CH or CY, which will be encoded again as follows:

Chain connections correspond to type CH, which is created for encoding either types DB, DC, and DD. Type CH represents a mutually connected subset of multiple relational element vertices, meaning any of its relational element vertices is mutually connected to any of the other relational element vertices within a set of multiple parallel arrowed lines indicating opposite directions.

For encoding element vertices of type CH, we denote the row labels as both c and b, and denote the column labels as both a and d. In the binary adjacency matrix, the entries of the row and column labels are defined as $z_{c,\,a} = z_{b,\,d} = 1$ for all $(b,\ d) \in Za$, where Za is a subset of all row and column labels in a set Z. The related vertices of these row and column labels are mutually connected. Assuming that the direction from the relational element vertex r_b to the relational element vertex r_d, and we focus on the starting relational element vertex r_b. Here, the relational element vertex r_b, where b should be transformed to $CH(b)$, is represented by $r_{CH(b)}$. Let us define $CH(b)$ as equal to a set of all $r_d \in R$ such that $z_{c,\,a} = z_{b,\,d} = 1$. Thus, type CH of a subset $r_{CH(b)}$, where b should be new sequentially obtained subset of the row labels b, and should be denoted again.

For type CY, if there are some relational element vertex numbers that are not included in the subset $CH(b)$ and those relational element vertices consistently have directed cycles either from or to the subset of the relational vertices $r_{CH(b)}$, then those vertices should be included in the subset $r_{CH(b)}$, which become type CY. As shown in Fig. 2, we call the paths between the relational element vertices and the subset of the relational element vertices $r_{CH(b)}$ as cyclic connections.

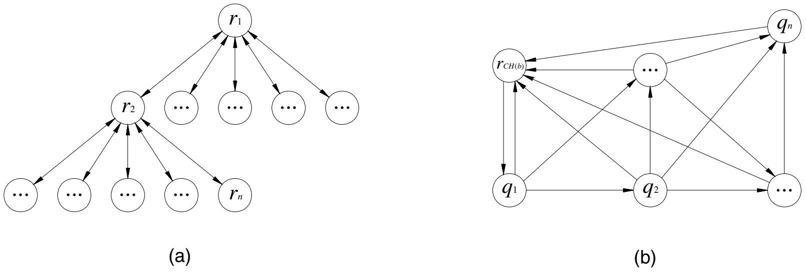


Fig. 2. Two different encoding types of connections: (a) chain connections and (b) cyclic connections

Let us denote these directed cycles, excluding the relational element vertices in the subset of the relational element vertices $r_{CH(b)}$, as *Zb*, which is a cyclic directed subset of all row and column labels in a set *Z*, and we finally denote a subset of relational element vertices as *CC*(*b*), where *b* is the number of each element vertex representing q_b, *b* = 1, 2,…, *n*. Here, for type CY, we denote a subset of all of the row and column labels as *Zc*, which should contain all relational element vertices of the subset *Za* and *Zb*. In addition, for numbering the relational element vertices, a subset of these numbers is denoted by *CY*(*b*), which should include all relational element vertices of the subset *CH*(*b*) and *CC*(*b*). Thus, type CY of a subset $r_{CY(b)}$, where *b* should be also new sequentially obtained subset of the row labels *b*, and should be also denoted again.

4.3 Encoding Types for Double-Encoded Substring

As shown in Fig. 3, types DA, DB, DC, and DD correspond to double-encoded substrings. Type DA comes from the remaining encoded substrings of type C, which should be already excluded both type CH and CY. For type DA, the row and column labels are denoted by *i* and *j* for encoding relational element vertices in a digraph, a subset of row and column labels is denoted by *Zd*, and the entries of the row and column labels are defined as

$$z_{i,\,j} = 1 \text{ for } i \text{ and } j = 1, 2, \cdots, n, i \neq j, \text{ and all}(i,\ j) \notin Zc \cup Ze \cup Zf \cup Zg\,. \tag{2}$$

This means that two specific relational element vertices in the digraph can be encoded as two unique sites for DNA experiments.

For type DB, the row and column labels are denoted by k_1 and k_2 for encoding relational element vertices in a digraph, a subset of row and column labels is denoted by *Ze*, and the entries of the row and column labels are defined as

$$z_{k_1,\,k_2} = 1 \text{ for } k_1 \text{ and } k_2 = 1, 2, \cdots, n, k_1 \neq k_2, \text{ and all}(k_1, k_2) \notin Zc \cup Zd \cup Zf \cup Zg\,. \tag{3}$$

This means that two subsets of specific relational element vertices in the digraph can be encoded as two unique sites for DNA experiments.

For type DC, the row and column labels are denoted by *i* and *k* for encoding relational element vertices in a digraph, a subset of row and column labels is denoted by *Zf*, and the entries of the row and column labels are defined as

$$z_{i,\,k} = 1 \text{ for } i = 1, 2, \cdots, n, k \in CY(b), i \neq k, \text{ and all}(i, k) \notin Zc \cup Zd \cup Ze \cup Zg\,. \tag{4}$$

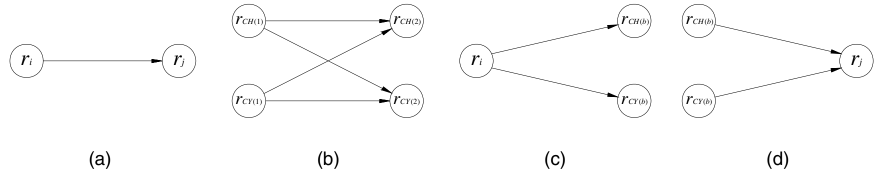


Fig. 3. Four different types of double-encoded substrings: (a) a direction from r_i to r_j, (b) directions from both $r_{CH(1)}$ and $r_{CY(1)}$ to both $r_{CH(2)}$ and $r_{CY(2)}$, (c) directions from r_i to both $r_{CH(b)}$ and $r_{CY(b)}$, and (d) directions from both $r_{CH(b)}$ and $r_{CY(b)}$ to r_j

This means that a specific relational vertex and a subset of specific relational element vertices in the digraph can be also encoded as two unique sites for DNA experiments.

For type DD, the row and column labels are denoted by k and j for encoding relational element vertices in a digraph, a subset of row and column labels is denoted by Zg, and the entries of the row and column labels are defined as

$$z_{k,\,j} = 1 \text{ for } k \in CY(b),\ j = 1, 2, \cdots, n,\ k \neq j, \text{ and all}(k,\ j) \notin Zc \cup Zd \cup Ze \cup Zf\ . \quad (5)$$

This means that a subset of specific relational element vertices and a specific relational vertex in the digraph can be also encoded as two unique sites for DNA experiments.

4.4 Encoding Types for Complementary Sites

As shown in Fig. 4, types DE and DF correspond to complementary sites. Type DE is basically constructed for attaching two different types of double-encoded substrings, corresponding to types DB and DB, types DB and DD, types DC and DB, and types DC and DD. Each of these two different types is sequentially lined up together in a 5' to 3' direction.

To attach each of the different double-encoded substrings, we must distinguish between circular or linear DNA fragments. The circular fragments correspond to either type CH or type CY. For this selection process, type DF is temporarily created; afterwards, type DF is recreated for attaching specific double-encoded substrings, corresponding to types DA and DA, types DA and DC, types DD and DA, and types DD and DC. Each of these two different types is also lined up together sequentially in a 5' to 3' direction.

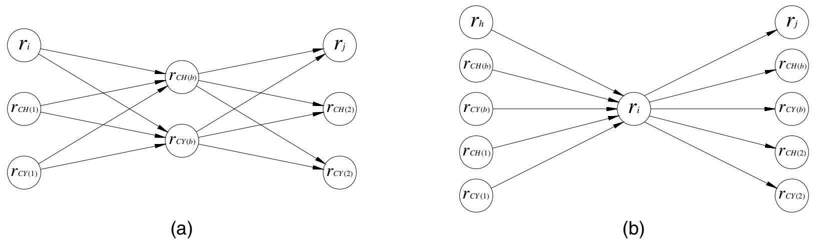


Fig. 4. Two different types of complementary sites: (a) complementary sites of types DB and DB, types DB and DD, types DC and DB, and types DC and DD, and (b) complementary sites of types DA and DA, types DA and DC, types DD and DA, and types DD and DC

5 Conclusions

A biologically intelligent method of encoding DNA sequences for the example of hierarchical structural modelling is proposed in this paper. This encoding method, applicable to different types of relational problems, shows the efficiency of employing molecular computation for a hierarchical classification of relational elements in a digraph.

In the example case presented, the proposed method encodes element vertices in a digraph, on the basis of the binary adjacency matrix and different encoding types. This then easily classifies a set of all relational element vertices into two or more hierarchical groups. In our future work, the proposed encoding method will be applied to methods other than hierarchical structural modelling, to tackle relational problems also composed of interconnected elements, which are connected to each other in relevant issues.

References

1. Warfield, J.N.: Structuring complex systems. In: Battelle Monograph, Columbus, Ohio, vol. (4) (1974)
2. Warfield, J.N.: On arranging elements of a hierarchy in graphic form. IEEE Transactions on Systems, Man, and Cybernetics SMC-3(2), 121–132 (1973)
3. Farris, G.F.: Executive decision making in organizations: Identifying the key men and managing the process, No. 551. MIT Sloan School Working Paper, 11–16 (1971)
4. Watada, J.: DNA computing and its application. In: Fulcher, J., Jain, L.C. (eds.) Computational Intelligence: A Compendium, pp. 1065–1086. Springer, Heidelberg (2008)
5. Rose, J.A., Hagiya, M., Deaton, R.J., Suyama, A.: A DNA-based in vitro Genetic Program. Journal of Biological Physics 28, 493–498 (2002)
6. Jeng, D.J.-F., Kim, I., Watada, J.: Bio-soft computing with fixed-length DNA to a group control optimization problem. Soft Computing 12, 223–228 (2006)
7. Kim, I., Jeng, D.J.-F., Watada, J.: Building a managerial support system for work rotation based on bio-soft computing. In: 2006 Annual Conference of the North American Fuzzy Information Processing Society, Proceedings, pp. 686–691 (2006)
8. Sakamoto, K., Kiga, D., Komiya, K., Gouzu, H., Yokoyama, S., Ikeda, S., Sugiyama, H., Hagiya, M.: State transitions by molecules. Biosystems 52, 81–91 (1999)
9. Adleman, L.: Molecular computation of solutions to combinatorial problems. Science 266, 1021–1024 (1994)
10. Hartl, D., Jones, E.: Essential genetics: A genomics perspective, 3rd edn., pp. 210–242. Jones and Bartlett Publishers, Inc. (2005)

A Bio-inspired Evolutionary Approach to Identifying Minimal Length Decision Rules in Emotional Usability Engineering

Ikno Kim and Junzo Watada

Graduate School of Information, Production and Systems
Waseda University
2-7 Hibikino, Wakamatsu-ku, Kitakyushu 808-0135, Japan
octoberkim@akane.waseda.jp, watada@waseda.jp

Abstract. Many of the applied methods and measurement tools of emotional usability engineering have been recommended for use designing products. A rough set method can also be a useful tool to be integrated with the basic concepts of emotional usability engineering. If such a method is applied, the groups of sensory words have to be investigated and their values reduced and classified to provide comprehensive information to product designers. However, a computational problem exists regarding the number of samples, groups of sensory words, and values required when resolving sense-based minimal decision rules. Considering this problem, we discuss the use of DNA computing, and propose a bio-inspired evolutionary method based on the rough set method, which should provide a new tool for emotional usability engineering.

1 Introduction

Accompanying the rapid development of technology, a wide variety of many different brands of complex products are now distributed through global markets and can be easily purchased. Because of this surplus, customers are increasingly inclined to focus more on the product design than on its functional capabilities. Emotional usability engineering (or *Kansei* engineering) is a novel method of measuring the feelings of existing and potential customers, and has recently been paid attention as part of the requirements for product designs [1].

In the emotional usability engineering process, points of views can be interpreted as having the ability to be classified into specific samples and groups of sensory words as data. A method of rough sets is a useful tool that can be easily adapted to this case and used for grasping characteristics of the classified samples and groups of sensory words [2].

Although a computable number of samples, groups of sensory words, and their values are often divided into sense-based minimal rules. But large numbers of samples and groups of sensory words often occur in the area of product design. In this paper, we propose a bio-inspired evolutionary method that answers this numerical problem in the emotional usability engineering process, and also show a novel way of using DNA molecules and biological techniques that provides a unique perspective on applied methods of emotional usability engineering. Further, we expand the capabilities

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 181–187, 2009.

of DNA computing, based on our previous works [3], [4], [5], to minimise the sense-based rules towards a biological tool of emotional usability engineering.

2 Emotional Usability Engineering with Rough Sets

A rough set method can be a powerful tool when dealing with making decisions related to impressions and feelings via emotional usability engineering. A lot of different kinds of samples composed of groups of sensory words or sentences with uncertain data are normally produced using the rough set method. In this section, the rough set method is described to help show our approach to identifying minimal length decision rules.

2.1 Rough Set Method

Rough set theory is a mathematical concept underlying set theory; the rough set method is based on the rough set theory. The rough set theory was proposed by Pawlak [6], who also produced several rough set theory studies [7], [8] and identified a number of advantages to evaluating significant data, minimising sets of decision rules, and analysing hidden data.

Let us assume that a set of samples is a universe denoted by U. A constraint set on samples is associated with a disjoint relation of classifying samples in emotional usability engineering. The constraint set on samples is denoted by T in each of samples x and y, and each is an element of a universe U, representing $(x, y)\in T$, but (x, y) is not equal to (y, x) and both (x, x) and (y, y) can each be elements of T. If there are three different elements x, y, and z, then the constraint set T satisfies the three relations that are reflexivity, symmetry, and transitivity. If a relation satisfies the above three properties, then the relation is called an equivalence relation.

In the rough set method, the equivalence relation is often called a discernibility relation. Here, all the samples in the universe are classified into sets of three disjoints, based on the discernibility relation [9]. The sets of these three disjoints are the main concept, and they are called the lower approximation, the upper approximation, and the boundary region.

2.2 Decision System

In the rough set method, an information system is different from a decision system, and therefore an information table is also different from a decision table. The main difference is that the information table is composed of a finite set of samples and attributes, whereas, the decision table is composed of three sets, two of which are the same as the information table's contents, and a third that is a set of decision attributes. For this paper, considering emotional usability engineering, (1) the attributes (the basis of the information table) are groups of sensory words; and (2) the decision attributes (the third set in the decision table) are brands. Table 1 shows an example decision table.

Table 1. Example of a decision table composed of 10 samples, 3 groups of sensory words, and 2 different brands

Sample	Group of Sensory Words-1	Group of Sensory Words-2	Group of Sensory Words-3	Brand
Sample-1	Traditional	Hot	Elegant	Brand B
Sample-2	Elastic	Cool	Sombre	Brand A
Sample-3	Soft	Cold	Fashionable	Brand B
Sample-4	Fixed	Tepid	Elegant	Brand B
Sample-5	Elastic	Warm	Casual	Brand A
Sample-6	Original	Cool	Formal	Brand A
Sample-7	Traditional	Tepid	Fashionable	Brand B
Sample-8	Fixed	Warm	Sombre	Brand A
Sample-9	Soft	Cold	Casual	Brand B
Sample-10	Original	Hot	Formal	Brand A

A decision system often deals with rough sets. The decision system is denoted by DS, and it is defined as $DS = (U, \Gamma, \varepsilon, \omega)$, where U is a finite set of samples, Γ is a finite set of groups of sensory words, ε is a brand, and ω is a value assignment. An element of a group of sensory words is denoted as ζ. Γ and is composed of n groups of sensory words $\zeta_1, \zeta_2, \ldots, \zeta_n$. Next, for dealing with values, let us denote a set of values of sensory word groups Γ^{c}_{ζ}. Assuming n values of sensory word groups exist for each group of sensory words, this is denoted by $\psi^{\zeta 1}_{1}, \psi^{\zeta 1}_{2}, \ldots, \psi^{\zeta 1}_{n}$ for $\Gamma^{c}_{\zeta 1}$, by $\psi^{\zeta 2}_{1}, \psi^{\zeta 2}_{2}, \ldots, \psi^{\zeta 2}_{n}$ for $\Gamma^{c}_{\zeta 2}$, $\cdots$, and $\psi^{\zeta n}_{1}, \psi^{\zeta n}_{2}, \ldots, \psi^{\zeta n}_{n}$ for $\Gamma^{c}_{\zeta n}$, respectively in the decision system. For decision values of brands, assuming that n decision values of the brand are denoted by $\tau_1, \tau_2, \ldots, \tau_n$, and a set of decision values is denoted by D^{v}.

3 Biological Computation Approach

A bio-inspired evolutionary method is specifically proposed as a new tool for emotional usability engineering to identify all minimal length decision rules. In this paper, we make the bio-inspired evolutionary method operate efficiently through DNA oligonucleotides on the basis of DNA computing. Computation with DNA molecules, called DNA computing, was firstly proposed by Adleman [10].

Both DNA oligonucleotides and DNA computing are associated with what is called Watson-Crick complementarity. This phenomenon is essential to the bio-inspired evolutionary method.

4 Bio-inspired Evolutionary Method

The main element in the bio-inspired evolutionary method is the encoding of DNA sequences on the basis of a binary adjacency matrix. Here, the binary adjacency matrix can be constructed from the directional relation composed of between samples and groups of sensory words, shown in Fig. 1. In this section, we describe the bio-inspired evolutionary method.

4.1 Directional Relation for Emotional Usability Engineering

Fig. 1 shows the directional relation for emotional usability engineering, in which there are two distinctive sorts - a group of samples and a group of pairs. Here, each pair consists of both a group of sensory words and a value of sensory word groups.

The example in Fig. 1 shows the directional relation of 10 samples and 3 groups of sensory words where each group has different values of sensory word groups. The 10 samples represent $U = \{x_1, x_2, \ldots, x_{10}\}$, and the 3 groups of sensory words represent ζ_1, ζ_2, and ζ_3. Each of the groups of sensory words has 5 values of sensory word groups, representing a pair matrix K that can be expressed as follows:

$$K = \begin{bmatrix} K(\zeta_1, \psi^{\zeta_1}{}_1) & K(\zeta_1, \psi^{\zeta_1}{}_2) & K(\zeta_1, \psi^{\zeta_1}{}_3) & K(\zeta_1, \psi^{\zeta_1}{}_4) & K(\zeta_1, \psi^{\zeta_1}{}_5) \\ K(\zeta_2, \psi^{\zeta_2}{}_1) & K(\zeta_2, \psi^{\zeta_2}{}_2) & K(\zeta_2, \psi^{\zeta_2}{}_3) & K(\zeta_2, \psi^{\zeta_2}{}_4) & K(\zeta_2, \psi^{\zeta_2}{}_5) \\ K(\zeta_3, \psi^{\zeta_3}{}_1) & K(\zeta_3, \psi^{\zeta_3}{}_2) & K(\zeta_3, \psi^{\zeta_3}{}_3) & K(\zeta_3, \psi^{\zeta_3}{}_4) & K(\zeta_3, \psi^{\zeta_3}{}_5) \end{bmatrix}. \quad (1)$$

This pair matrix is mainly used for encoding double-encoded substrings for both type 3 and type 6 in a DNA encoding part.

For representations of directions between both (1) samples and pairs; and (2) pairs and another pairs. Here, the three types of relational directions are defined as (1) when a sample x_i is directed to a pair (ζ_α, $\psi^{\zeta\alpha}{}_\alpha$); (2) when a pair ($\zeta_\beta$, $\psi^{\zeta\beta}{}_\beta$) is directed to a sample x_j; and (3) when a pair (ζ_α, $\psi^{\zeta\alpha}{}_\alpha$) is directed to a pair (ζ_β, $\psi^{\zeta\beta}{}_\beta$). A subset of the directions from the sample x_i to the pair (ζ_α, $\psi^{\zeta\alpha}{}_\alpha$) is denoted as B^a, a subset of the directions from the pair (ζ_β, $\psi^{\zeta\beta}{}_\beta$) to the sample x_j is denoted as B^b, and a subset of the directions from the pair (ζ_α, $\psi^{\zeta\alpha}{}_\alpha$) to the pair ($\zeta_\beta$, $\psi^{\zeta\beta}{}_\beta$) is denoted as B^c. A set of the subsets B^a, B^b, and B^c is denoted by B, and this set is composed of the entire existent directions and represents $B = B^a \cup B^b \cup B^c$.

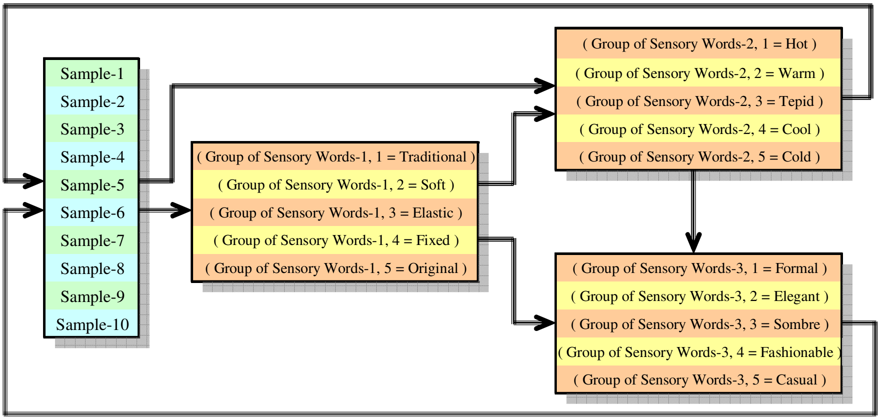

Sample-1
Sample-2
Sample-3
Sample-4
Sample-5
Sample-6
Sample-7
Sample-8
Sample-9
Sample-10
(Group of Sensory Words-1, 1 = Traditional)
(Group of Sensory Words-1, 2 = Soft)
(Group of Sensory Words-1, 3 = Elastic)
(Group of Sensory Words-1, 4 = Fixed)
(Group of Sensory Words-1, 5 = Original)
(Group of Sensory Words-2, 1 = Hot)
(Group of Sensory Words-2, 2 = Warm)
(Group of Sensory Words-2, 3 = Tepid)
(Group of Sensory Words-2, 4 = Cool)
(Group of Sensory Words-2, 5 = Cold)
(Group of Sensory Words-3, 1 = Formal)
(Group of Sensory Words-3, 2 = Elegant)
(Group of Sensory Words-3, 3 = Sombre)
(Group of Sensory Words-3, 4 = Fashionable)
(Group of Sensory Words-3, 5 = Casual)

Fig. 1. Directional relations between samples and groups of sensory words for emotional usability engineering

4.2 Encoding Samples and Pairs

First, type 1 corresponds to a double-encoded substring, which indicates each of all two different single elements that are the sample x_i and the pair $(\zeta_\alpha, \psi^{\zeta\alpha}{}_{\alpha})$. For type 1, these two different types of samples have a directional relation indicating the direction from the sample x_i to the pair $(\zeta_\alpha, \psi^{\zeta\alpha}{}_{\alpha})$.

Second, type 2 also corresponds to a double-encoded substring, which indicates each of all two different elements that are the pair $(\zeta_\beta, \psi^{\zeta\beta}{}_{\beta})$ and the sample x_j. For type 2, these two different types of samples also have a directional relation indicating the direction from the pair $(\zeta_\beta, \psi^{\zeta\beta}{}_{\beta})$ to the sample x_j.

Third, type 3 also corresponds to a double-encoded substring, which indicates each of all two different elements that are the pair $(\zeta_\alpha, \psi^{\zeta\alpha}{}_{\alpha})$ and the other pair $(\zeta_\beta, \psi^{\zeta\beta}{}_{\beta})$. For type 3, these two different types of pairs also have a directional relation indicating the direction from the pair $(\zeta_\alpha, \psi^{\zeta\alpha}{}_{\alpha})$ to the other pair $(\zeta_\beta, \psi^{\zeta\beta}{}_{\beta})$.

Finally, all types 4, 5, 6, and 7 correspond to complementary sites that are between two of the double-encoded substrings, which are (1) types 1 and 3; (2) types 3 and 2; (3) types 3 and 3; and (4) types 2 and 1. The forces of the hybridisation and ligation techniques make all three different substrings connect to each other.

5 Experimental Procedure

For experiments, each complementary site is described by each concatenation of DNA fragments, which correspond to complementary sites of both samples and groups of sensory words. All of the substrings for encoding double-encoded substrings and their complementary sites should be generated.

The fitting restriction enzymes should have been added to each sample of the substrings and complementary sites in this encoding process. The function of restriction enzymes is to split DNA fragments composed of a loop [11], [12]. All the circular DNA fragments are split and become linear DNA sequences. The procedure is as follows:

First, as shown in Fig. 1, a directional relation for emotional usability engineering is constructed by the directional relations between the given samples, groups of sensory words, and values of sensory word groups. The double-encoded substrings and the complementary sites are constructed on the basis of encoding the directed samples and pairs. The existent directions of the pairs in types 3 and 6 are first constructed and encoded to determine subsets of the lower approximations in each brand.

Second, all of the encoded pairs in type 3 and its complementary sites in type 6 are artificially synthesized and loaded into a test tube. For this hybridisation, the DNA sequences and their complementary sites are hybridised on the basis of the Watson-Crick complementary rules. Use of electrophoresis apparatus makes DNA fragments separate according to the size of the hybridised DNA fragments. The size of the divided DNA strands is measured using electrophoresis apparatus. If there are two or more hybridised DNA sequences of the same length that means those hybridised DNA sequences are composed of samples that have exactly the same values of sensory word groups.

Third, the hybridised DNA sequences in types 3 and 6 are denatured to become DNA sequences in single-stranded DNA, which will be reused for the second step of the hybridisation and ligation process. We identify some samples that only belong to one of the lower approximation subsets and are associated with the same brand while only connecting together to a specific pair. This process reduces encoding DNA sequences, and the reduced sequences are used to identify minimal length decision rules in emotional usability engineering.

Fourth, types 1, 2, 4, 5, and 7 are constructed second and encoded, reusing the encoded DNA sequences of types 3 and 6 to identify all the minimal length decision rules in each brand. All of the encoded samples and pairs in types 1 and 2 and its complementary sites in types 4, 5, and 7, as well as the added pairs of type 3 and type 6 are artificially synthesized and loaded into a test tube.

Fifth, as shown in Fig. 2, all the circular DNA sequences are identified and distinguished from the entire hybridised DNA sequences with ligations. The circular type of DNA sequences are split at any one point by restriction of enzymes to become linear sequences, before classifying these into each group of samples.

Finally, all of each sample group of the split DNA sequences are measured, and each group of samples is classified into each test tube. After this, if there are two or more DNA sequences that are the same length, they are removed. However, if two or more DNA sequences correspond to two or more samples that are from one of lower approximation subsets, then they are not removed. Each sample of the remaining DNA sequences in each brand contains two or more pairs, which are distinguished by

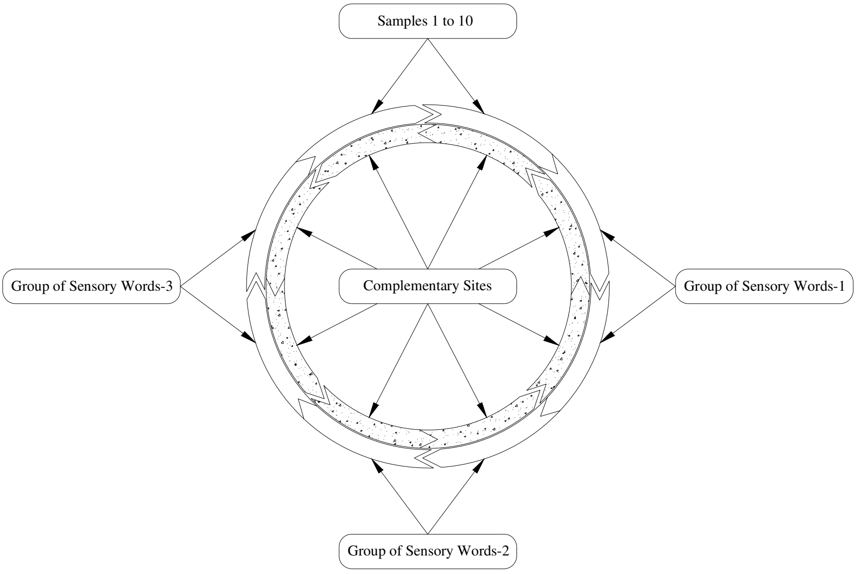


Fig. 2. The example representation of a circular DNA sequence for emotional usability engineering

an affinity separation method. A subset of each sample in each lower approximation has its own subset of pairs, which is marked to become a group of decision rules in each brand.

6 Conclusions

A bio-inspired evolutionary method has been proposed in this paper for the minimization of the sense-based decision rules. The minimised rules have been resolved, on the basis of the several discoveries technique, for both specific encodings and lengths of selected DNA sequences.

Subsets of the lower approximations have also been classified using the proposed process of several biological techniques. We have shown that this bio-inspired evolutionary method, composed mainly of the two above processes, can be employed by product designers, especially those with a huge number of samples corresponding to sense-based information.

References

1. Kanda, T.: Colors and Kansei. In: 2004 IEEE International Conference on Systems, Man, and Cybernetics, pp. 299–305 (2004)
2. Kobayashi, H., Ota, S.: The semantic network of Kansei words. In: 2000 IEEE International Conference on Systems, Man, and Cybernetics, pp. 690–694 (2000)
3. Watada, J.: DNA computing and its application. In: Fulcher, J., Jain, L.C. (eds.) Computational Intelligence: A Compendium, pp. 1065–1086. Springer, Heidelberg (2008)
4. Jeng, D.J.-F., Kim, I., Watada, J.: Bio-soft computing with fixed-length DNA to a group control optimization problem. In: Soft Computing, vol. 12, pp. 223–228. Springer, Heidelberg (2006)
5. Kim, I., Jeng, D.J.-F., Watada, J.: Building a managerial support system for work rotation based on bio-soft computing. In: 2006 Annual Conference of the North American Fuzzy Information Processing Society, Proceedings, pp. 686–691 (2006)
6. Pawlak, Z.: Rough sets. International Journal of Computer and Information Sciences 11(5), 341–356 (1982)
7. Pawlak, Z., Wong, S.K.M., Ziarko, W.: Rough sets: Probabilistic versus deterministic approach. International Journal of Man-Machine Studies 29(1), 81–95 (1988)
8. Grzymala-Busse, J.W.: A comparison of tree strategies to rule induction from data with numerical attributes. Electronic Notes in Theoretical Computer Science 82(4), 132–140 (2003)
9. Peters, J.F., Skowron, A.: Transactions on Rough Sets VI. LNCS, vol. 4374, pp. 351–396. Springer, Heidelberg (2007)
10. Adleman, L.: Molecular computation of solutions to combinatorial problems. Science 266, 1021–1024 (1994)
11. Watson, J.D., Myers, R.M., Caudy, A.A., Witkowski, J.A.: Recombinant DNA, 3rd edn., pp. 79–106. W. H. Freeman and Company, New York (2007)
12. Hartl, D., Jones, E.: Essential genetics: A genomics perspective, 3rd edn., pp. 210–242. Jones and Bartlett Publishers, Inc. (2005)

Determining Workstation Groups in a Fixed Factory Facility Based on Biological Computation

Ikno Kim and Junzo Watada

Graduate School of Information, Production and Systems
Waseda University
2-7 Hibikino, Wakamatsu-ku, Kitakyushu 808-0135, Japan
octoberkim@akane.waseda.jp, watada@waseda.jp

Abstract. A strategy for making layout decisions is an important element in developing operating systems in manufacturing factories or other industrial plants. In this paper, we look at fixed factory facilities and propose a method for designing different sorts of layouts related to factories running at high-volume and producing a low-variety of products. Where many tasks are called, each with a different task time, it can be difficult to arrange a fixed factory facility in the optimal way. Therefore, we propose a computational method using DNA molecules for designing production systems by determining all the feasible workstation groups in a fixed factory facility, and we show that this computation method can be generally applied to layout decisions.

1 Introduction

Layout design is important, particularly for manufacturing factories, and this has implications for current business strategies. The main purpose of the design is to develop a new and cost-effective layout that meets competitive needs. Careful layout design in fixed factory facilities can deliver a long period of efficient operation; it can also increase company assets in terms of quality of work life, operational capacity, product quality, and so forth, and thereby increase the competitive efficiency of the whole operation [1].

Layout decisions are particularly important for any manufacturing factory that has plans either to develop existing facilities or establish new ones. In this context, the product-focused facility can also be referred to in terms of the fixed factory facility that handles an undiversified, high-volume production, in which the products are of similar types or have similar features. In this paper, we focus specifically on detecting productive groups of tasks in the fixed factory facility to determine the most practical workstations where there are fixed-position machines and where there is limited plant space.

Although it is not difficult to identify a feasible layout where production involves a limited number of tasks, it is difficult to find productive groups of tasks and determine all the feasible workstation positions when a large number of tasks are involved. We therefore propose a new biological computation method that can find productive groups of tasks and determine feasible workstation situations, helping to construct an efficient layout. This new method can be applied to regenerating strategies in operations management.

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 188–194, 2009.

2 Tasks in Graph Theory

In order to determine workstation positions in a fixed factory facility, in which there are several given tasks, each with a task time, all the tasks can be transformed to a node; task times can be transformed to a weighted node. From focusing on tasks in graph theory, we can determine all groups of practical workstations by finding the largest time length of tasks in a clique. In this section, we briefly describe the maximum number of tasks and the largest time length of tasks in a clique.

2.1 Maximum Number of Tasks in a Clique

A fixed factory facility is composed of tasks, task times, and possible routing lines between the given tasks. In the fixed factory facility, tasks are labelled M_i, i = 1, 2,..., n, and each task has its own time, which is denoted by t_i, i = 1, 2,..., n. Fig. 1 shows an example of a fixed factory facility composed of 12 tasks with task times.

For the tasks without times, let us denote a graph of the fixed factory facility G = (M, R), where M is a task set of G. A possible routing line of distinct tasks in M represents $R \subseteq M \times M$, which is a set of all the possible routing lines of G. In the graph of the fixed factory facility, a clique Q is a subset of tasks, and if $G(Q)$ is complete, then there is a possible routing line between every pair of distinct tasks. The number of tasks in a clique is defined as $\eta(G) = \max\{|Q_s| : Q_s$ is a set of tasks in a clique in $G\}$, where Q_s is the cardinality of a set, and the number of tasks contained in Q_s denoted by $|Q_s|$. Here, the main purpose of the maximum number of tasks in a clique is to determine tasks in a clique that is the maximum cardinality in the graph of the fixed factory facility.

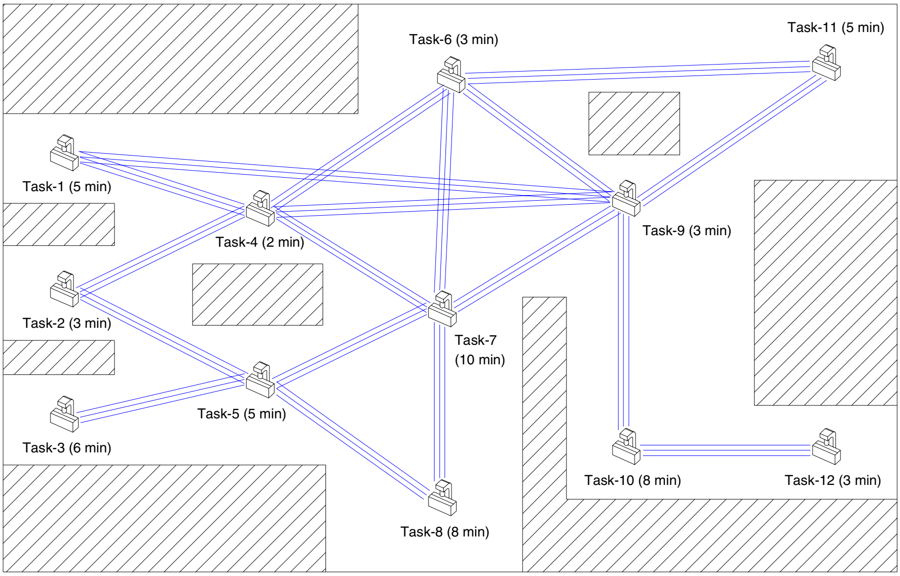


Fig. 1. Example of a fixed factory facility composed of 12 tasks with their possible routings

2.2 Largest Time Length of Tasks in a Clique

A graph of the fixed factory facility with tasks and task times is denoted as $G = (M, R, \varepsilon)$, where ε is a mapping that is composed of sets of disjoints between both tasks and possible routing lines together with this mapping. Some tasks of the possible routing lines can be unordered pairs representing $(i, j)\in M^2$, where i is not equal to j.

The mapping with task times is denoted as ε, which assigns each possible routing line representing $(i, j)\in R$. A positive task time here is denoted as both $\varepsilon(i, i) > 0$ and $\varepsilon(i, j) = \varepsilon(j, i) > 0$. Also, this mapping with task times assigns each task $i\in M$, a positive task time, denoted by t_i, which is associated with i, where a set of the graph of the fixed factory facility with task times is ordered by n.

The main purpose of determining the largest time length of tasks in a clique is concerned with the total task time, which is summed by each task time of the maximum tasks in a clique for the fixed factory facility. The largest time length of tasks in a clique is defined as $\eta(G, t) = \max\{W(Q_s) : Q_s$ is a set of task times in a clique in $G\}$.

The largest time length of tasks in a clique is formulated in an integer programming formulation. For all of $(i, j)\in R_q$, where $R_q = \{(i, j) \mid i, j\in M,\ i \neq j$ and (i, j) is not an element of $R\}$, the quadratic expressions $x_i x_j = 0$ since for $x_i, x_j\in\{0, 1\}$ and $x_i + x_j \leq 1$, if and only if $x_i x_j = 0$ makes use of the constraints for the minimisation problem converted from the problem of the maximum number of tasks in a clique [2]. Adding the quadratic terms twice makes the constraints remove, and these terms reach the objective function representing

$$f(x) = -\sum_{i=1}^{n} x_i + 2 \sum_{(i,j)\in R_q,\, i>j} x_i x_j = x^T (L_{Gq} - I)x\,, \tag{1}$$

where x^T corresponds to a transpose of a vector x, L_{Gq} corresponds a symmetric matrix in a complement graph G, and I corresponds to the identity matrix. The quadratic terms can be represented as $x_i x_j = 0$, which means that the largest time length of tasks in a clique representing

$$\text{minimise } f(x) = x^T Lx\,, \tag{2}$$

$$\text{subject to } x\in \{0,\ 1\}^n\,, \tag{3}$$

where L is an adjacency matrix of G_q, where $G_q = (M, R_q)$. The specific set $Q_s = \lambda(x^*)$ is the largest time length of tasks in a clique in the fixed factory facility with the task time $W(Q_s)$ that is equal to $-f(x^*)$, if an optimal solution is denoted by x^* that is an optimal solution to the above minimum problem.

3 Computational Method with DNA Molecules

The given tasks in the graph of the fixed factory facility can be shown to be groups of feasible workstations by identifying the largest time length of tasks in a clique. For this case, in this section, we describe a computational method with DNA molecules that arranges all of tasks with task times, use the results of this arrangement, and determine all the feasible workstation groups.

3.1 Biological Computation

A computational method with DNA molecules is executed with the functions of DNA molecules. The main function is referred to a phenomenon, which is called Watson-Crick complementarity [3]. DNA strands are composed of the four nucleotides adenine, guanine, cytosine and thymine. Watson-Crick complementarity corresponds to two main reactions that are the bond between adenine and thymine, and the bond between guanine and cytosine.

The computational method with DNA molecules can be called biological computation, because the features of biological molecules are used for computation. A computation with DNA molecules was first proposed by Adleman [4]. Based on our previous works [5], [6], [7], we show a new biological method with DNA molecules that determines the largest time length of tasks in a clique, as well as arranging all tasks composed of task times in order of time length. Each group of these orders can be used for determining the position of feasible workstations in the fixed factory facility.

$$\begin{array}{l}
\text{Task-1} \\
\text{Task-2} \\
\text{Task-3} \\
\text{Task-4} \\
\text{Task-5} \\
\text{Task-6} \\
\text{Task-7} \\
\text{Task-8} \\
\text{Task-9} \\
\text{Task-10} \\
\text{Task-11} \\
\text{Task-12}
\end{array}
\left[\begin{array}{cccccccccccc}
1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\
0 & 1 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 0 & 1 & 0 \\
0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1
\end{array}\right]$$

Fig. 2. Connectivity matrix of the model fixed factory facility

3.2 Encoding Tasks and Task Times

The best way of representing each fragment of DNA sequences is with a connectivity matrix. The connectivity matrix can be constructed by the relations between tasks in the fixed factory facility.

In such a matrix, ordering rows and columns reveals the weighted tasks, corresponding to tasks with task times. Fig. 2 shows an example of the connectivity matrix, with the rows and columns composed of 12 tasks with each task time. The entry of the matrix is denoted as $b_{i,j}$, i and j = 1, 2,..., n, which gives the value of the connection from task i to task j, and records which pairs of tasks are adjacent to each other. In Fig. 2, if two specific tasks M_i and M_j are adjacent, then the entry $b_{i,j}$ should be 1; whereas if two specific tasks, M_i and $M_{j,}$ are not adjacent, then the entry $b_{i,j}$ should be 0. The relationship of these binary numbers make the best use of this proposed computational method when DNA molecules are used to resolve how many DNA fragments will be needed for the biological experiments. The computational method with DNA molecules arranges all of tasks with task times in the fixed factory facility in the following manner.

First, if the task is included in either a clique or a set of two tasks, it is set to the value 1, otherwise it is set to the value 0. In the case of a fixed factory facility, the entry value 1 corresponds to one of the 12 tasks included in either tasks in a clique or a set of two tasks. The entry value 0 also corresponds to one of the 12 tasks, but one that is specifically not included in other tasks. Based on this, all the possible combinations between the given tasks with possible routing lines should be created.

Second, the graph of the fixed factory facility contains all possible routing lines, which are missing in the original graph. All task pairs connected in a complementary relationship should be removed using restriction enzymes. For the fixed factory facility example, the remaining pool selects the DNA sequences from the value 1 to 12 for the given 12 tasks.

Third, gel electrophoresis apparatus can be used to find specific tasks by measuring the length of the DNA strands. The shortest length strand corresponds to the maximum number of tasks in a clique, and the second shortest length strand corresponds to the second largest number of tasks in a clique. The above procedure is repeated until all the tasks are detected.

Finally, all the tasks in each clique should be arranged and the remaining sets of paired tasks that are connected should be in only one possible routing line. Also, they should be marked in order of time length. The arrangement of all the tasks in each clique can be used to determine the position of all the feasible workstation groups.

4 Simulated Biological Experiments

Each task of all the DNA sequences consists of two different sequences, which are called both a routing sequence denoted by RS_k, k = 1, 2,..., n, and a task sequence MS_i, i = 1, 2,..., n. With regard to the synthesised DNA fragments, the two routing sequences are created to connect a task of the DNA sequence, and the task sequence is created to distinguish whether the two routing sequences contain that task or not.

To design each task of the DNA sequence, for the example of the fixed factory facility, we should consider (1) the binary numbers in the connectivity matrix; and (2) the possible routing lines between the 12 tasks with each task time. If all of the 12 tasks are not included in both the tasks in a clique and a set of two tasks, this means the entry values of the 12 tasks are 0. For the 12 tasks with each task time, all of the lengths of the DNA sequence are the same; if the task has the longer task time, then the task sequence is longer in the DNA sequence.

The length of DNA base pairs for each routing sequence RS_k is set by the time length of the tasks, so the longer the time, the shorter the length of base pairs. The length of DNA base pairs for each task sequence MS_i is also set by the time length of the tasks, but it in reverse of the base pairs - the longer time corresponds to the longer length of base pairs.

The impossible routing lines of the complementary sequence should be removed using restriction enzyme sites. To remove all of the impossible routing lines, the specific 12 restriction enzyme sites can slit the impossible routing lines of the DNA sequence. The split DNA sequences cannot be amplified by the polymerase chain reactions.

To measure the length of DNA base pairs, we use gel electrophoresis apparatus, which can classify the DNA sequences in terms of size. This allows us to determine the length of the classified DNA strands, and compare them to the other measured DNA strands. All of the weighted tasks in each clique can then be properly classified by task time and the number of cliques, using this technique.

5 Simulated Experimental Results

All of the tasks in each clique can be determined according to the measured length of the DNA sequence, and we can make the resulting graph of the fixed factory facility, which includes either sets of tasks in each clique or sets of paired tasks. The following results show the sets of tasks in each clique and the sets of paired tasks.

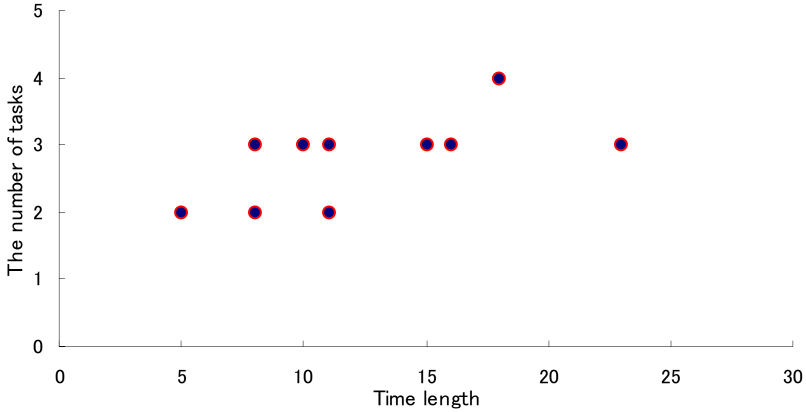


Fig. 3. Comparison of time length and the number of tasks

The largest time length of three tasks in a clique is composed of the tasks M_5, M_7, and M_8. The second largest time length of four tasks in a clique is composed of the tasks M_4, M_6, M_7, and M_9. The third largest time length of three tasks is composed of the tasks M_6, M_7, and M_9. The two fourth largest time lengths of three tasks are composed of (1) the tasks M_4, M_6, and M_7; and (2) the tasks M_4, M_7, and M_9. The fifth largest time length of three tasks in a clique is composed of the tasks M_6, M_9, and M_{11}. The sixth largest time length of three tasks in a clique is composed of the tasks M_1, M_4, and M_9. The smallest time length of three tasks in a clique is composed of the tasks M_4, M_6, and M_9. Also, there are three different time lengths of paired task sets detected. Based on these results, we created Fig. 3, which shows a comparison of time length and the number of tasks, which can be used to place workstations within a fixed factory facility.

6 Conclusions

In this paper, a new method of biological computation is proposed for three purposes: to detect productive groups of tasks, determine feasible workstation positions, and show that this method might be adapted for making layout decisions where a large number of tasks are involved. The simulated experiment represented each task with its own task-time.

In this study, we have mentioned the importance of layout decisions in manufacturing and industrial factories or plants, for operating strategies. The most advantageous element of dealing with DNA molecules in executing this computation is that it provides enormous scope for determining the specific tasks necessary to a process and for solving complex problems of this kind as they emerge in production.

References

1. Heizer, J., Render, B.: Operations management, 9th edn., pp. 356–370. Pearson Education, Inc., London (2008)
2. Bomze, I.M., Budinich, M., Pardalos, P.M., Pelillo, M.: The maximum clique problem. In: Handbook of Combinatorial Optimization, pp. 3–47. Kluwer Academic Publishers, Dordrecht (1999)
3. Watson, J.D., Myers, R.M., Caudy, A.A., Witkowski, J.A.: Recombinant DNA, 3rd edn., pp. 79–106. W. H. Freeman and Company, New York (2007)
4. Adleman, L.: Molecular computation of solutions to combinatorial problems. Science 266, 1021–1024 (1994)
5. Watada, J.: DNA computing and its application. In: Fulcher, J., Jain, L.C. (eds.) Computational Intelligence: A Compendium, pp. 1065–1086. Springer, Heidelberg (2008)
6. Jeng, D.J.-F., Kim, I., Watada, J.: Bio-soft computing with fixed-length DNA to a group control optimization problem. In: Soft Computing, vol. 12, pp. 223–228. Springer, Heidelberg (2006)
7. Kim, I., Jeng, D.J.-F., Watada, J.: Building a managerial support system for work rotation based on bio-soft computing. In: 2006 Annual Conference of the North American Fuzzy Information Processing Society, Proceedings, pp. 686–691 (2006)

A Fuzzy Risk Assessment in Software Development Defuzzified by Signed Distance

Huey-Ming Lee[1] and Lily Lin[2]

[1] Department of Information Management, Chinese Culture University
55, Hwa-Kung Road, Yang-Ming-San, Taipei (11114), Taiwan
hmlee@faculty.pccu.edu.tw
[2] Department of International Business, China University of Technology
56, Sec. 3, Hsing-Lung Road, Taipei (116), Taiwan
lily@cute.edu.tw

Abstract. In this paper, we present computational rule inferences to tackle the rate of aggregative risk in fuzzy circumstances. Based on the maximum membership grade principle, we apply the signed distance to defuzzify which is better than by the centroid. The proposed fuzzy assessment method is easier, closer to evaluator real thinking and more useful than the ones which have presented before.

Keywords: Risk assessment; fuzzy risk assessment.

1 Introduction

Generally, risk is the traditional manner of expressing uncertainty in the systems life cycle. Risk assessment is a common first step and also the most important step in a risk management process. Risk assessment is the determination of quantitative or qualitative value of risk related to a concrete situation and a recognized threat. In a quantitative sense, it is the probability at such a given point in a system's life cycle that predicted goals can not be achieved with the available resources. Due to the complexity of risk factors and the compounding uncertainty associated with future sources of risk, risk is normally not treated with mathematical rigor during the early life cycle phases [1]. Risks result in project problems such as schedule and cost overrun, so risk minimization is a very important project management activity [12]. Up to now, there are many papers investigating risk identification, risk analysis, risk priority, and risk management planning [1-4, 6-7].

Based on [2-4, 6-7], Lee [9] classified the risk factors into six attributes, divided each attribute into some risk items, and built up the hierarchical structured model of aggregative risk and the evaluating procedure of structured model, ranged the grade of risk for each risk item into eleven ranks, and proposed the procedure to evaluate the rate of aggregative risk using two stages fuzzy assessment method. Chen [5] ranged the grade of risk for each risk item into thirteen ranks, and defuzzified the trapezoid or triangular fuzzy numbers by the median. Lee and Lin [10] proposed a new fuzzy assessment method to tackle the rate of aggregative risk in fuzzy circumstances.

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 195–202, 2009.

Based on the maximum membership grade principle, we apply the signed distance to defuzzify which is better than by the centroid [11].

2 The Proposed Fuzzy Risk Assessment Method

We present the fuzzy assessment method as follows;

Step 1: Assessment form for the risk items:

The criteria ratings of risk are linguistic variables with linguistic values V_1, V_2, …, V_7, where V_1 = extra low, V_2= very low, V_3= low, V_4= middle, V_5= high, V_6= very high, V_7= extra high. These linguistic values are treated as fuzzy numbers with triangular membership functions as follows:

$$\tilde{V}_1 = (0, 0, \frac{1}{6}),$$
$$\tilde{V}_k = (\frac{k-2}{6}, \frac{k-1}{6}, \frac{k}{6}),\ for\ k = 2, 3, \ldots, 6 \tag{1}$$
$$\tilde{V}_7 = (\frac{5}{6}, 1, 1)$$

In previous studies [5, 9], the evaluator only chooses one grade from grade of risk for each risk item, it ignores the evaluator's incomplete and uncertain thinking. Therefore, if we use fuzzy numbers of assessment in fuzzy sense to express the degree of evaluator's feelings based on his own concepts, the computing results will be closer to the evaluator's real thought.

The assessment for each risk item with fuzzy number can reduce the degree of subjectivity of the evaluator, express the degree of evaluator's feelings based on his own concepts. The results will be closer to the evaluator's real thought. Based on the structured model of aggregative risk proposed by Lee and Lin [10] and evaluating form of structured model proposed by Lee [9], we propose the assessment form of the structured model as shown in Table 1 and propose a new assessment method using computational rule inference to tackle the rate of aggregative risk in software development.

In Table 1,

$$\sum_{i=1}^{6} W_2(i) = 1, \quad 0 \le W_2(i) \le 1\text{,} \tag{2}$$

for each i = 1, 2, ... 6.

Table 1. Contents of the hierarchical structure model [10]

Attribute	Risk item	Weight-2	Weight-1	Linguistic variables						
				V_1	V_2	V_3	V_4	V_5	V_6	V_7
X_1: Personal		$W_2(1)$								
	X_{11}: Personal shortfalls, key person(s) quit		$W_1(1,1)$	$m_{11}^{(1)}$	$m_{11}^{(2)}$	$m_{11}^{(3)}$	$m_{11}^{(4)}$	$m_{11}^{(5)}$	$m_{11}^{(6)}$	$m_{11}^{(7)}$
X_2: System requirement		$W_2(2)$								
	X_{21}: Requirement ambiguity		$W_1(2,1)$	$m_{21}^{(1)}$	$m_{21}^{(2)}$	$m_{21}^{(3)}$	$m_{21}^{(4)}$	$m_{21}^{(5)}$	$m_{21}^{(6)}$	$m_{21}^{(7)}$
	X_{22}: Developing the wrong software function		$W_1(2,2)$	$m_{22}^{(1)}$	$m_{22}^{(2)}$	$m_{22}^{(3)}$	$m_{22}^{(4)}$	$m_{22}^{(5)}$	$m_{22}^{(6)}$	$m_{22}^{(7)}$
	X_{23}: Developing the wrong user interface		$W_1(2,3)$	$m_{23}^{(1)}$	$m_{23}^{(2)}$	$m_{23}^{(3)}$	$m_{23}^{(4)}$	$m_{23}^{(5)}$	$m_{23}^{(6)}$	$m_{23}^{(7)}$
	X_{24}: Continuing stream requirement changes		$W_1(2,4)$	$m_{24}^{(1)}$	$m_{24}^{(2)}$	$m_{24}^{(3)}$	$m_{24}^{(4)}$	$m_{24}^{(5)}$	$m_{24}^{(6)}$	$m_{24}^{(7)}$
X_3: Schedules and budgets		$W_2(3)$								
	X_{31}: Schedule not accurate		$W_1(3,1)$	$m_{31}^{(1)}$	$m_{31}^{(2)}$	$m_{31}^{(3)}$	$m_{31}^{(4)}$	$m_{31}^{(5)}$	$m_{31}^{(6)}$	$m_{31}^{(7)}$
	X_{32}: Budget not sufficient		$W_1(3,2)$	$m_{32}^{(1)}$	$m_{32}^{(2)}$	$m_{32}^{(3)}$	$m_{32}^{(4)}$	$m_{32}^{(5)}$	$m_{32}^{(6)}$	$m_{32}^{(7)}$
X_4: Developing technology		$W_2(4)$								
	X_{41}: Gold-plating		$W_1(4,1)$	$m_{41}^{(1)}$	$m_{41}^{(2)}$	$m_{41}^{(3)}$	$m_{41}^{(4)}$	$m_{41}^{(5)}$	$m_{41}^{(6)}$	$m_{41}^{(7)}$
	X_{42}: Skill levels inadequate		$W_1(4,2)$	$m_{42}^{(1)}$	$m_{42}^{(2)}$	$m_{42}^{(3)}$	$m_{42}^{(4)}$	$m_{42}^{(5)}$	$m_{42}^{(6)}$	$m_{42}^{(7)}$
	X_{43}: Straining hardware		$W_1(4,3)$	$m_{43}^{(1)}$	$m_{43}^{(2)}$	$m_{43}^{(3)}$	$m_{43}^{(4)}$	$m_{43}^{(5)}$	$m_{43}^{(6)}$	$m_{43}^{(7)}$
	X_{44}: Straining software		$W_1(4,4)$	$m_{44}^{(1)}$	$m_{44}^{(2)}$	$m_{44}^{(3)}$	$m_{44}^{(4)}$	$m_{44}^{(5)}$	$m_{44}^{(6)}$	$m_{44}^{(7)}$
X_5: External resource		$W_2(5)$								
	X_{51}: Shortfalls in externally furnished components		$W_1(5,1)$	$m_{51}^{(1)}$	$m_{51}^{(2)}$	$m_{51}^{(3)}$	$m_{51}^{(4)}$	$m_{51}^{(5)}$	$m_{51}^{(6)}$	$m_{51}^{(7)}$
	X_{52}: Shortfalls in externally performed tasks		$W_1(5,2)$	$m_{52}^{(1)}$	$m_{52}^{(2)}$	$m_{52}^{(3)}$	$m_{52}^{(4)}$	$m_{52}^{(5)}$	$m_{52}^{(6)}$	$m_{52}^{(7)}$
X_6.: Performance		$W_2(6)$								
	X_{61}: Real-time performance shortfalls		$W_1(6,1)$	$m_{61}^{(1)}$	$m_{61}^{(2)}$	$m_{61}^{(3)}$	$m_{61}^{(4)}$	$m_{61}^{(5)}$	$m_{61}^{(6)}$	$m_{61}^{(7)}$

$$\sum_{i=1}^{n_k} W_1(k,i) = 1,\ \ 0 \le W_1(k,i) \le 1 \tag{3}$$

for k=1, 2,..., 6; $n_1 = 1, n_2 = 4, n_3 = 2, n_4 = 4, n_5 = 2, n_6 = 1.$; i = 1, 2,..., n_k.

$$\sum_{l=1}^{7} m_{ki}^{(l)} = 1, 0 \le m_{ki}^{(l)} \le 1, \tag{4}$$

for l=1, 2, ..., 7; k=1, 2, ..., 6; i=1, 2, …, n_k.

From Table 1, we directly use the fuzzy numbers ($m_{ki}^{(l)}$) rather than the linguistic values to evaluate. Also, we may express the risk item X_{ki} as fuzzy discrete type

$$X_{ki} = \frac{m_{ki}^{(1)}}{V_1} \oplus \frac{m_{ki}^{(2)}}{V_2} \oplus \frac{m_{ki}^{(3)}}{V_3} \oplus \frac{m_{ki}^{(4)}}{V_4} \oplus \frac{m_{ki}^{(5)}}{V_5} \oplus \frac{m_{ki}^{(6)}}{V_6} \oplus \frac{m_{ki}^{(7)}}{V_7} \tag{5}$$

Step 2: Weighted triangular fuzzy numbers

For easy to express, we take some one attribute, saying X_j, and the items, saying X_{j1}, X_{j2}, …, X_{jn_j} in Table 1, and introduce the weighted triangular fuzzy numbers as shown in Table 2, for j=1, 2, …, 6, and $n_1 = 1, n_2 = 4, n_3 = 2, n_4 = 4, n_5 = 2, n_6 = 1.$

Table 2. Contents of the weighted triangular fuzzy number for item X_{jk}

Attribute	Risk item	Linguistic variables						
		V$_1$	V$_2$	V$_3$	V$_4$	V$_5$	V$_6$	V$_7$
X_j								
	X_{j1}	$m_{j1}^{(1)}$	$m_{j1}^{(2)}$	$m_{j1}^{(3)}$	$m_{j1}^{(4)}$	$m_{j1}^{(5)}$	$m_{j1}^{(6)}$	$m_{j1}^{(7)}$
Weighted triangular fuzzy number		$m_{j1}^{(1)}\tilde{V}_1$	$m_{j1}^{(2)}\tilde{V}_2$	$m_{j1}^{(3)}\tilde{V}_3$	$m_{j1}^{(4)}\tilde{V}_4$	$m_{j1}^{(5)}\tilde{V}_5$	$m_{j1}^{(6)}\tilde{V}_6$	$m_{j1}^{(7)}\tilde{V}_7$
	X_{j2}	$m_{j2}^{(1)}$	$m_{j2}^{(2)}$	$m_{j2}^{(3)}$	$m_{j2}^{(4)}$	$m_{j2}^{(5)}$	$m_{j2}^{(6)}$	$m_{j2}^{(7)}$
Weighted triangular fuzzy number		$m_{j2}^{(1)}\tilde{V}_1$	$m_{j2}^{(2)}\tilde{V}_2$	$m_{j2}^{(3)}\tilde{V}_3$	$m_{j2}^{(4)}\tilde{V}_4$	$m_{j2}^{(5)}\tilde{V}_5$	$m_{j2}^{(6)}\tilde{V}_6$	$m_{j2}^{(7)}\tilde{V}_7$
	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.
	X_{jn_j}	$m_{jn_j}^{(1)}$	$m_{jn_j}^{(2)}$	$m_{jn_j}^{(3)}$	$m_{jn_j}^{(4)}$	$m_{jn_j}^{(5)}$	$m_{jn_j}^{(6)}$	$m_{jn_j}^{(7)}$
Weighted triangular fuzzy number		$m_{jn_j}^{(1)}\tilde{V}_1$	$m_{jn_j}^{(2)}\tilde{V}_2$	$m_{jn_j}^{(3)}\tilde{V}_3$	$m_{jn_j}^{(4)}\tilde{V}_4$	$m_{jn_j}^{(5)}\tilde{V}_5$	$m_{jn_j}^{(6)}\tilde{V}_6$	$m_{jn_j}^{(7)}\tilde{V}_7$

Let $B = \{\tilde{V}_1, \tilde{V}_2, ..., \tilde{V}_7\}$. From Table 2, we can form the fuzzy relation on X_j and B with weighted triangular fuzzy number elements as follows:

$$\tilde{R}_j = \begin{bmatrix} m_{j1}^{(1)}\tilde{V}_1 & m_{j1}^{(2)}\tilde{V}_2 & \dots & m_{j1}^{(7)}\tilde{V}_7 \\ m_{j2}^{(1)}\tilde{V}_1 & m_{j2}^{(2)}\tilde{V}_2 & \dots & m_{j2}^{(7)}\tilde{V}_7 \\ . & & & \\ . & & & \\ . & & & \\ m_{jn_j}^{(1)}\tilde{V}_1 & m_{jn_j}^{(2)}\tilde{V}_2 & \dots & m_{jn_j}^{(7)}\tilde{V}_7 \end{bmatrix} \tag{6}$$

Step 3: The first stage computational rule of inference

We let

$$(\tilde{T}_{j1}, \tilde{T}_{j2}, ..., \tilde{T}_{j7}) = (w_1(j,1), w_1(j,2), ..., w_1(j,n_j) \bullet \tilde{R}_j \tag{7}$$

where

$$\tilde{T}_{jq} = w_1(j,1) m_{j1}^{(q)} \tilde{V}_q \oplus w_1(j,2) m_{j2}^{(q)} \tilde{V}_q \oplus \ ... \oplus w_1(j,n_j) m_{jn_j}^{(q)} \tilde{V}_q \tag{8}$$

for j=1, 2, …, 6; q=1, 2, …, 7. We have that $\tilde{T}_{jq}$ is a triangular fuzzy number.

Defuzzified $\tilde{T}_{jq}$ by the signed distance [13], we have

$$d(\tilde{T}_{jq}, \tilde{0}) = \sum_{t=1}^{n_j} w_1(j,t) \cdot m_{jt}^{(q)} \cdot d(\tilde{V}_q, \tilde{0}) \tag{9}$$

is rate of risk for the attribute X_j.

Step 4: The second stage computational rule of inference

We let

$$(\tilde{S}_1, \tilde{S}_2, ..., \tilde{S}_7)$$

$$= (w_2(1), w_2(2), ..., w_2(6)) \bullet \begin{bmatrix} \tilde{T}_{11} \ \tilde{T}_{12} \ ... \ \tilde{T}_{17} \\ \tilde{T}_{21} \ \tilde{T}_{22} \ ... \ \tilde{T}_{27} \\ . \\ . \\ . \\ \tilde{T}_{61} \ \tilde{T}_{62} \ ... \ \tilde{T}_{67} \end{bmatrix} \tag{10}$$

where

$$\tilde{S}_q = w_2(1)\tilde{T}_{1q} \oplus w_2(2)\tilde{T}_{2q} \oplus ... \oplus w_2(6)\tilde{T}_{6q} \tag{11}$$

for q=1, 2, …, 7.

Defuzzified $\tilde{S}_q$ by the signed distance [13], we have

$$d(\tilde{S}_q, \tilde{0}) = \sum_{t=1}^{6} w_2(t) \cdot d(\tilde{T}_{tq}, \tilde{0}) \tag{12}$$

is aggregative rate of risk.

Then we have the following Proposition 1.

Proposition 1 Let $S_{jq}=d(\tilde{T}_{jq},\tilde{0})$, then, we have

(1) for the attribute X_j, and rating risk V_q, the rate of risk is S_{jq}

(2) the rate of risk for the attribute X_j is $Q^{(j)}=\sum_{q=1}^{7}S_{jq}$. (13)

(3) let T_q be the defuzzified of $\tilde{S}_q$ by signed distance, then, the rate of risk for the rating risk V_q is

$$T_q=\sum_{j=1}^{6}w_2(j)\cdot d(\tilde{T}_{jq},\tilde{0}) \tag{14}$$

(4) the aggregative rate of risk is

$$Rate=\sum_{q=1}^{7}T_q \tag{15}$$

Table 3. Contents of the example [10]

Attribute	Risk item	Weight-2	Weight-1	Linguistic variables						
				V1	V2	V3	V4	V5	V6	V7
X_1		0.3								
	X_{11}		1	0	0.17	0.83	0	0	0	0
X_2		0.3								
	X_{21}		0.4	0	0.53	0.47	0	0	0	0
	X_{22}		0.4	0	0.89	0.11	0	0	0	0
	X_{23}		0.1	0.25	0.75	0	0	0	0	0
	X_{24}		0.1	0.61	0.39	0	0	0	0	0
X_3		0.1								
	X_{31}		0.5	0	0.17	0.83	0	0	0	0
	X_{32}		0.5	0	0.53	0.47	0	0	0	0
X_4		0.1								
	X_{41}		0.3	0	0.89	0.11	0	0	0	0
	X_{42}		0.1	0	0.17	0.83	0	0	0	0
	X_{43}		0.3	0	0.17	0.83	0	0	0	0
	X_{44}		0.3	0	0.53	0.47	0	0	0	0
X_5		0.1								
	X_{51}		0.5	0	0	0.81	0.19	0	0	0
	X_{52}		0.5	0	0	0.81	0.19	0	0	0
X_6		0.1								
	X_{61}		1	0	0.17	0.83	0	0	0	0

3 Numerical Example

Example: Assume that we have the following attributes, weights, grade of risk for each risk item as shown in Table 3 [10].

(1) By the Proposition 1 shown in Section 2, we have the following computed results.

	V_1	V_2	V_3	V_4	V_5	V_6	V_7	$Q^{(j)}$
X_1	0	0.028333	0.276667	0	0	0	0	0.305
X_2	0.003583	0.113667	0.077333	0	0	0	0	0.194583
X_3	0	0.058333	0.216667	0	0	0	0	0.275
X_4	0	0.082333	0.168667	0	0	0	0	0.251
X_5	0	0	0.27	0.095	0	0	0	0.365
X_6	0	0.028333	0.276667	0	0	0	0	0.305
T_q	0.001075	0.0595	0.1994	0.0095	0	0	0	

(2) the aggregative rate of risk is $Rate = 0.269475$

(3) Comparison with Lin and Lee [10]

a) In [10], the rate of aggregative risk is 0.269647. By the proposed method in this study, the computed result is 0.269475. The relative error is (0.269475-0.269647)/0.269647= -0.00064. It is very small. But, the proposed method is easier than in [10].

b) We can tackle the risk rate of each attribute by the proposed method in this study.

4 Conclusion

In general survey forces evaluator to assess one grade from the grade of risk to each risk item, but it ignores the uncertainty of human thought. For instance, when the evaluator need to choose the assessment from the survey which lists eleven choices including “definitely unimportant”, “extra unimportant”, “very unimportant”, “unimportant”, “slightly unimportant”, “middle”, “slightly important”, “important”, “very important”, “extra important”, and “definitely important”, the general survey becomes quiet exclusive. The assessment of evaluation with fuzzy numbers can reduce the degree of subjectivity of the evaluator. Based on the maximum membership grade principle, if the fuzzy number of $\underset{\sim}{A}$ is not an isosceles triangle, then defuzzified $\underset{\sim}{A}$ by the signed distance method is better than by the centroid method. In this paper, we applied the signed distance method to defuzzify to evaluate the rate of aggregative risk.

Acknowledgments

The authors gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

References

1. AFSC: Software Risk Abatement, U. S. Air Force Systems Command, AFSC/AFLC pamphlet 800-45, Andrews AFB, MD, pp. 1–28 (september 1988)
2. Boehm, B.W.: Software Risk Management. CS Press, Los Alamitos (1989)
3. Boehm, B.W.: Software Risk Management: Principles and Practices. IEEE Software 8, 32–41 (1991)
4. Charette, R.N.: Software Engineering Risk Analysis and Management. Mc. Graw-Hill, New York (1989)
5. Chen, S.M.: Evaluating the Rate of Aggregative Risk in Software Development Using Fuzzy Set Theory. Cybernetics and Systems: International Journal 30, 57–75 (1999)
6. Conger, S.A.: The New Software Engineering. Wadsworth Publishing Co, Belmont (1994)
7. Gilb, T.: Principles of Software Engineering Management. Addison-Wesley Publishing Co., New York (1988)
8. Kaufmann, A., Gupta, M.M.: Introduction to Fuzzy Arithmetic Theory and Applications. Van Nostrand Reinhold, New York (1991)
9. Lee, H.-M.: Applying fuzzy set theory to evaluate the rate of aggregative risk in software development. Fuzzy sets and Systems 79, 323–336 (1996)
10. Lee, H.-M., Lin, L.: A New Fuzzy Risk Assessment Approach. In: Lovrek, I., Howlett, R.J., Jain, L.C. (eds.) KES 2008, Part III. LNCS (LNAI), vol. 5179, pp. 98–105. Springer, Heidelberg (2008)
11. Lee, H.-M., Lin, L.: Fuzzy Facility Site Selection Model Based on Signed Distance Method, International Journal of Innovative Computing. Information and Control 5(6), 1505–1514 (2009)
12. Sommerville, I.: Software Engineering, 6th edn. Pearson Education Limited, England (2001)
13. Yao, J.-S., Wu, K.: Ranking Fuzzy Numbers Based on Decomposition Principle and Signed Distance. Fuzzy Sets and Systems 116, 275–288 (2000)

Particle Swarm Optimization for Multi-function Worker Assignment Problem

Shamshul Bahar Yaakob and Junzo Watada

Graduate School of Information, Production and Systems, 2-7, Hibikino, Wakamatsu,
Kitakyushu, 808-0135 Japan
shamshul@fuji.waseda.jp

Abstract. A problem of worker assignment in cellular manufacturing (CM) environment is studied in this paper. The worker assignment problem is an NP-complete problem. In this paper, worker assignment method is modeled based on the principles of particle swarm optimization (PSO). PSO applies a collaborative population-based search, which models over the social behavior of fish schooling and bird flocking. PSO system combines local search method through self-experience with global search methods through neighboring experience, attempting to balance the exploration-exploitation trade-off which determines the efficiency and accuracy of an optimization. An effect of velocity controlled for the PSO's is newly included in this paper. We applied the adaptation and implementation of the PSO search strategy to the worker assignment problem. Typical application examples are also presented: the results demonstrate that the velocity information is an important factor for searching best solution and our method is a viable approach for the worker assignment problem.

Keywords: worker assignment, particle swarm optimization, cellular manufacturing.

1 Introduction

A recent trend shows increasing popularity of Cellular Manufacturing (CM), and other team-related approaches in the workspace to achieve the goal of lean manufacturing and to effectively solve mass-production [1], [2]. Here, the CM is defined as the grouping of people and process, or machines, into specific areas dedicated to the production of a family of parts [3]. A lot of study/research has been performed to improve the grouping of machines and parts into cells as a result of this trend towards CM. CM that has been performed solely on machine-part interaction has frequently shown limited benefits [4]. This failure has lead researchers to search for other factors that influence the performance of CM, culminating in an increasing interest in the effect of personal skills and quality in the performance of teams [5]. Therefore, the concept of the flexibility of workers comes into consideration [6]. The possibility that workers may help each other and share their workloads are closely relevance with the multi-function worker that is able to perform a given subset of types of tasks. Hence, the enhancing feelings of interpersonal justice and equity need to evaluate [7], [8], [9].

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 203–211, 2009.

PSO is an algorithm that follows a collaborative population-based search model. Each individual of the population, called a ‘particle’, flies around in a multidimensional search space looking for the optimal solution [10] Particles, then, may adjust their position according to their own and their neighboring-particles experience, moving toward their best position or their neighbor’s best position. In order to achieve this, particles keep previously reached ‘best’ positions in a cognitive memory. PSO performance measured according to a predefined fitness function [11]. Balancing between global and local exploration abilities of the flying particles could be achieved through user-defined parameters. PSO has many advantages over other heuristic techniques such that it can be implemented in a few lines of computer code, it requires only primitive mathematical operators, and it has great capability of escaping local minima.

The objective of this paper is to find the optimal workers combination of work assignment in CM. The effect of relationship evaluation on multi-function workers’ in a CM is considered, and a worker assignment model is established with an objective of maximizing the total grade; which is, the proposed method will assign which worker to which machine of which cell. We proposed a new method of worker assignment, which is based on the PSO method. Then, the performance of the proposed method is compare with other existing methods under various worker assignments in CM environment.

The remainder of the paper is organized as follows. Section 2 defines the worker assignment problem. Section 3 describes our PSO algorithm for the worker assignment problem. Results of numerical example are reported in Section 4. Finally, Section 5 concludes the paper.

2 Multi-function Worker Assignment Problem

In this section, the evaluation including the relationship among the workers is developed in order to tackle the worker assignment problems efficiently [12]. The evaluation criteria are classified into three factors: social factors, performance factors, and mental factors.

In this paper the relationship is evaluated pairwise among workers. The relationship evaluation is performed via the sum of all the evaluation results between any couple of workers. This prescription can be generally applied to any size of the worker group, and is appropriate for our purpose and for computations. In general each criterion has its importance weight depending on the nature of jobs. Therefore, in the following computation method the weighted sum is performed (see Eq. (2)). Suppose the following situation: the decision makers are responsible for assessing the suitability of m workers (P_i, $i = 1,\ldots, m$) under each of the k criteria (C_t, $t = 1,\ldots k$). Let $e(J, i, C_t)$ be a number, which is a rating assigned to a worker P_i by the decision makers for a criterion (C_t) for a job (J). Let $W(J, C_t)$ be the importance weight of the criterion C_t for the job J. The decision makers can fix the total worker number assigned to each job depending on the job feature, if required. If not, the total the number of workers is also determined in our algorithm. The computation flow of our method is as follows:

Step 1: Determine the evaluation criteria. Select the appropriate rating scale to assess the importance weights of the criteria and the suitability of the workers to the criteria. Tabulate suitability ratings (*S*) assigned to each worker (*P*) for each criterion (C_t) by each decision maker. Tabulate importance weightings ($W(J, C_t)$) assigned to each criterion (C_t) for each job (*J*) by the decision makers.

Step 2: A suitability ranking of each worker P_i for the job *J* can be obtained by standard arithmetic operations:

$$E_{eval}(J,i) = \frac{1}{k}\sum_{t=1}^{k} e(J,i,Ct)W(J,Ct). \tag{1}$$

In Eq. (1) the summation result is divided by the total number *k* of criteria employed so that $E_{eval}(J, i)$ does not depend on *k*. The ranking order is determined by the total grade value $E_{eval}(J, i)$ for each job *J*.

Step 3: In order to find possible combinations PCs, the decision makers assign the minimum grade value required for each job.

Step 4: Based on the result of the workers' suitability evaluation, the possible combinations PCs are obtained in order of the ranking each worker having a larger grade value is selected and assigned to the PC. The total grade value for the possible combination $E_{PC}(J)$ for the job *J* is as follows:

$$E_{PC}(J) = \sum_{i}^{PC} E_{eval}(J,i), \tag{2}$$

where the summation is performed over all members of the PC. The results are listed in the ranking order based on the total grade value $E_{PC}(J)$ for the possible combinations for each job. If any PC does not satisfy the minimum grade value, return to Step 3.

Step 5: Evaluate the relationship among the workers in a combination for a job: the relationship among the workers is computed as follows:

$$E_{RL}(J) = W_{RL}(J)\left[\frac{1}{{}_fC_2}\sum_{(i,j)} e_{RL}(i,j)\right], \tag{3}$$

where ${}_fC_2 = f(f- 1)/2$, $e_{RL}(i, j)$ is a value assigned to the relationship between two workers (P_i and P_j) in the combination, the summation is taken over all couples of workers in the job *J*, $W_{RL}(J)$ is the importance weight of the relationship for the job *J*, and $E_{RL}(J)$ is the total relationship-evaluation value.

The summation is normalized by ${}_fC_2$ so that $E_{RL}(J)$ does not depend on the number (${}_fC_2$) of couples of workers. The final evaluation is computed as follows:

$$E_{comb}(J) = E_{PC}(J) + E_{RL}(J). \tag{4}$$

The result for final evaluation is listed in the ranking order based on the grade value $E_{comb}(J)$.

Step 6: If the total job number *TJ* is one, the maximum grade value combination is the best one. When the total job number *TJ* is more than one, the decision makers specify if one worker can be assigned to plural jobs or not, depending on the job nature. If one worker is not assigned to plural jobs, an overlapped assignment of one worker is checked and avoided in the total combination construction. Based on this information, the total final combination evaluation TE_{comb} is as follows:

$$TE_{comb} = \sum_{J=1}^{TJ} E_{comb}(J). \tag{5}$$

The result for the total final evaluation TE_{comb} is listed. The combination that has the maximum grade value is the optimal solution for the worker assignment problem.

3 Particle Swarm Optimization and Multi-function Worker Assignment

3.1 Particle Swarm Optimization (PSO)

PSO was first introduced by Kennedy and Eberhart in 1995 and partly inspired by the behavior of large animal swarms such as schooling fish or flocking birds [13], [14]. PSO conducts search using a population of random solutions, corresponding to individual. In addition, each potential solution called particles is also assigned a randomized velocity. Each particle in PSO flies in the hyperspace with a velocity which is dynamically adjusted its position according to their own and their neighboring-particles experiences, moving towards two points: the best position so far by itself called P_{best} and by neighbor called G_{best} at every iteration. The particle swarm optimization concept consists of, at each time step, changing the velocity of each particle towards its P_{best} and G_{best}.

Suppose that the search space is *D* dimensional, then the i^{th} particle of the swarm can be represented by a *D* dimensional $X_i=(x_{i1}, x_{i2},\ldots,x_{iD})'$. The particle velocity can be represented by another *D* dimensional vector $V_i=(v_{i1},v_{i2},\ldots,v_{iD})'$. The best previously visited position of the i^{th} particle is denoted as $P_i=(p_{i1}, p_{i2},\ldots,p_{iD})'$. Defining *g* as the index of the best particle in the swarm, and let the superscripts denote the iteration number, then the position of a particle and its velocity are updated by the following equations

$$v_{id}^{k+1} = wv_{id}^{k} + c_1 r_1^k (p_{id}^k - x_{id}^k) + c_2 r_2^k (p_{gd}^k - x_{id}^k) \tag{6}$$

$$x_{id}^{k+1} = x_{id}^{k} + v_{id}^{k+1} \tag{7}$$

where $d=1,...,D$, $i=1,...,N$, and N is the size of swarm; w is called inertia weight; c_1, c_2 are two positive constants, called cognitive and social parameter respectively; r_1, r_2 are random number, uniformly distributed in [0,1]; and $k=1,2,...$ determines the iteration number.

3.2 Particle Swarm Optimization for Multi-function Worker Assignment Problem

In this section, we describe the formulation of a PSO algorithm for the worker assignment problem. In this paper, we set up a search space of M dimension of the worker assignment problem. Each dimension has discrete set of possible values limited to $s=\{P_i: 1<i<N\}$; such that N is the number of workers in the one cell in CM.

Using such particle representation, the PSO population is represented as two-dimensional array consisting of N particles, each represented as a vector of M job. Thus, a particle flies in an M-dimensional search space. A job is internally represented as an integer value indicating the worker number to which this job is assigned to during the course of PSO.

In PSO, the exploration-exploitation trade-off determines the efficiency and accuracy of an optimization, which are controlled by the velocity update equation [15]. Thus, the velocity information is important for searching best solution, and if we can control the velocity in different range, we could find the different best solution. In our proposed method, we discover that the velocity smaller, easier to find the best solution. Therefore, for our problem the following limit is applied; at first, if particle velocity is greater than v_{max}, it was set equal to v_{max}; if the velocity is smaller than v_{min}, it was set equal to v_{min}. Secondly, if the found solution is greater than p_{max}, it was set equal to p_{max}, if the found solution is smaller than p_{min}, it was set equal to p_{min}, where p_{min}, p_{max} is the range of solution.

In this research, M-job assignment is map into corresponding M-coordinate particle position. The algorithm starts by generating randomly as many as potential assignments for the problem as the size of the initial population of the PSO. It then measures particles' fitness. Eq. (4) is use as our fitness function.

The algorithm keeps an updated version of "global-best" position and "local-best" position through out the course of its execution. It does that by conducting two ongoing comparisons: First, it compares the fitness of each particle being in its 'current' position with fitness of other particles in the population in order to determine the global-best position each generation. Then, in order to determine a local-best position for every particle, it compares different visited positions of particles with its current position. These two positions affect the new velocity of every particle in the population according to Eq.(6). As shown in Eq.(6), two random parameters control amount of effect the two positions such as global and local best positions impose over the new particle velocity. Introducing a randomize and unbiased affect from either positions is the purpose of these parameters. Therefore, introduces a bit of exploration at some times and a bit of exploitation at another time randomly. The algorithm uses the new velocity to update the particle current position to a new position according to Eq.(7). Once all particles adjust their positions, they will constitute the new status of the PSO population. Then, the algorithm evaluates the fitness of these particles according to their new positions. Finally, the algorithm repeats this whole process of determining the global and the local best positions, updating particle position and evaluating new particle position until a user-determined criterion is satisfied. In our case, this criterion is mapped to a maximum number of generations. The proposed PSO algorithm for worker assignment problem is summarized in Formulation 1.

[Formulation 1]

```
Defines:
    $N$, number of workers given by the problem
    $M$, number of job
    $P$, size of the PSO population
    $PSO[i]$, position of the $i^{th}$ particle of the PSO population represented as
    an $M$-dimension vector, whose entries' values belong to the set $\{1,\ldots, N\}$;
    $PSO[i][j]$, worker number to which the $j^{th}$ job in the $i^{th}$ particle is assigned.
    Fitness $[i]$, grade-value function of the $i^{th}$ particle according to Eq. 4.
    $V[i]$, travelled distance (or velocity) of an $i^{th}$ particle represented as an
    $M$-dimensional real-coded vector.
    $G_{best}$ , an index to global-best position.
    $P_{best}[i]$, position of the local-best position of the $i^{th}$ particle.
    $P_{best\_fitness}\ [i]$, local best fitness for the best position visited by the $i^{th}$ particle.
For each particle $i$ in the population;
    For each job $j$;
    Initialize $PSO[i][j]$ randomly from the set $\{1,\ldots,N\}$.
    Initialize $V[i]$ randomly.
    Evaluate fitness$[i]$.
    Initialize $G_{best}$ with the index of the particle with the best fitness (highest
    grade-value) among the population
    Initialize $P_{best}\ [i]$ with a copy of $PSO[i]$ for All $i \le P$
Repeat until a number of generation equal to twice of the total job, is passed;
  ·   Find G$_{best}$ such that fitness $[G_{best}] \ge$ fitness$[i]$ for All $i \le P$
  ·   For each particle $i$; $PSO[i]$ if fitness$[i]>P_{best\_fitness}[i]$ for All $i \le P$
  ·   For each particle $i$; update $V[i]$ according to Eq. 6; if particle velocity is greater than $v_{max}$, it was set equal to $v_{max}$; if the velocity is smaller than $v_{min}$, it was set equal to $v_{min}$. Secondly, if the found solution is greater than $p_{max}$, it was set equal to $p_{max}$, if the found solution is smaller than $p_{min}$, it was set equal to $p_{min}$, where $p_{min}$, $p_{max}$ is the range of solution and $PSO\ [i]$ according to Eq. 7; if the found solution is greater than $p_{max}$, it was set equal to $p_{max}$, if the found solution is smaller than $p_{min}$, it was set equal to $p_{min}$, where $p_{min}$, $p_{max}$ is the range of solution.
  ·   Evaluate fitness$[i]$ for All $i \le P$
```

4 Numerical Example

It is assumed that a CM is composed of three cells. Each cell is provided with six machines which possess different functions, but the three cells are as same as each other. Several kinds of productions will be manufactured, the productive quantity of each production is given, and each production needs to be produced on which machine is known according to the process flow of each production itself. Twenty workers need to be assigning to the three cells, and each worker needs to be assigned to only one machine of only one cell. In this case, the possible combination from ${}_{20}C_6$; it becomes 38760 combinatorial.

In the proposed method, the PSO algorithm begins by initializing randomly a population of particles based on the given cell. Each entry in a particle array represents a job in the cell, and each value for these entries is a worker number to which the corresponding job is assigned. Initially, particle velocities are initialized randomly. To simplify the discussion, assume a population of ten particles initialized. As an example, in the first generation particle 1's position {1,4,5,2,6,3} suggest a solution to problem assigning job 1 to worker 1, job 2 to worker 4, job 3 to worker 5, and job 4 to worker 2, job 5 to worker 6, and job 6 to worker 3. Particle fitness, which is the grade-value of combination (calculated according to Eq. (4)), is shown adjacent to each particle. In this population, the best particle (G_{best}) has a fitness value. A copy of G_{best} position is updated throughout generations.

While it is the initial population, historical local best position for a PSO particle is its initial position. PSO then updates each particle velocity using Eq. (6). Then the updated velocities are used in Eq. (7) to get the new position for each particle. Since, values in a particle are worker numbers; a real value is meaningless. Therefore, in the algorithm we usually round these numbers to the closest worker number by dropping the sign of the fractional part. The effects of converting from continuous domain to discrete domain need to be determined further, which mapping will yield the better results. The new particle position is a new solution to the assignment problem. In conjunction, this adjustment is done for all particles and the cycle of finding the G_{best} and local bests and updating the particle positions continues until satisfactory results or termination criteria are met such as population converge onto one solution.

We used in experiments the following values for parameter that control each algorithm: the size of population equals to twice the number of jobs; the inertia weight w is set to 0.90; c_1=c_2=1.49; v_{max}=0.40, v_{min}= -0.40. Furthermore, for each experimental setting, 10 trials were performed.

Table 1. Summary of results comparison

Cell No.	Standard PSO method	Proposed method	Improvement in grading value (%)
1	*W4(M1),W7(M2),W8(M3), W11(M4),W14(M5),W15(M6)*	*W4(M1),W8(M2), W11(M3), W15(M4),W3(M5),W1(M6)*	2.50
2	*W1(M1),W3(M2),W5(M3), W12(M4),W16(M5),W18(M6)*	*W18(M1),W5(M2),W12(M3), W9(M4),W17(M5),W2(M6)*	0.20
3	*W2(M1),W6(M2),W9(M3), W10(M4),W13(M5),W17(M6)*	*W6(M1),W10(M2),W7(M3), W14(M4),W16(M5),W13(M6)*	5.29

**W, M represent worker and machine accordingly.*

Table 2. PSO algorithm evaluation results

	Standard PSO	PSO with controlled velocity
Goal Achieved	80%	100%
Iteration No.	418	252
Average Iteration No.	471	299
Best Value of Fitness	48.1	49.8
Average Best Value of Fitness	47.77	49.03

By using the method presented in this paper, the result is presented in Table 1. While standard PSO does not concern with the traveling velocity, the method proposed in this paper had limit the traveling velocity. Taking Cell 1 as a case, grade value for the best combination processed by standard PSO ((W4, M1), (W7, M2), (W8, M3), (W11, M4), (W14, M5), (W15, M6)) is 57.9. By using proposed method, the grade value 59.1 for the best combination ((W4, M1), (W8, M2), (W11, M3), (W15, M4), (W3, M5), (W1, M6)) as in Table 1 is higher than the total grade value 57.9 by 2.5%. In Cell 2, the combination ((W1, M1), (W3, M2), (W5, M3), (W12, M4), (W16, M5), (W18, M6)) is the highest grade value for the standard PSO method. By using proposed method, the grade value 58.2 for the best combination ((W18, M1), (W5, M2), (W12, M3), (W9, M4), (W17, M5), (W2, M6)). Finally for Cell 3, ((W2, M1), (W6, M2), (W9, M3), (W10, M4), (W13, M5), (W17, M6)) had resulted in the highest grade value for both standard PSO method and for the proposed method is 59.8 for the best combination ((W6, M1), (W10, M2), (W17, M3), (W14, M4), (W16, M5), (W13, M6)).

In order to evaluate the performance of the proposed PSO with controlled velocity, five evaluation criteria were selected in order to evaluate the performance of the PSO algorithm for worker assignment problem. The details are shown in Table 2. From the result, it is clear that both standard and proposed methods can find optimal solution for the problem. However, PSO algorithm with controlled velocity is faster and possess better precision for the optimal solution.

5 Conclusions

A new proposal was presented to solve the problem of worker assignment in CM environment. In this paper also a new heuristic algorithm called PSO algorithm is proposed for the worker assignment problem. The proposed method was used to typical application examples in CM environment. The results show that the proposed method is one of the key issues and our proposal is effective for the decision-making process.

These results also indicate that the proposed PSO algorithm is an attractive alternative for solving worker assignment problem. The method proposed in this paper is concentrated on particles PSO algorithm that suits discrete optimization problems.

Moreover, PSO application to other combinatorial optimization problems needs to be study in the future.

The domain of worker assignment illustrated in this paper is focused on a CM environment. It may also be applied to other types of assignment problems. The result also showed the grade values are significantly improved compared with a standard PSO method. While it is generally accepted that standard PSO approach is suitable to solve this kind of problem, result in this paper suggests that controlled velocity PSO is more superior than the standard method.

Acknowledgments. The first author would like to thank Universiti Malaysia Perlis and the Ministry of Higher Education Malaysia for a study leave in Waseda University under the SLAI-KPT scholarship.

References

1. John, P.W.: Worker allocation in lean U-shaped production lines. International Journal of Production Research 46(13), 3485–3502 (2008)
2. Li, Q., Gong, J., Tang, J., Song, J.: Simulation of the model of workers' assignment in cellular manufacturing based on the multifunctional workers. In: Chinese Control and Decision Conference, pp. 992–996 (2008)
3. Slomp, J., Bokhorst, J.A.C., Molleman, E.: Cross-training in a cellular manufacturing environment. Computer and Industrial engineering 48(3), 609–624 (2005)
4. Fitzpatrick, E.L., Askin, R.G.: Forming effective worker teams with multi-functional skill reqiurements. Journal of Computer & Industrial Engineering 48, 539–608 (2005)
5. Stefan, T., Monika, W.: Toward the measure of virtual teams effectiveness. Human Factors and Ergonomics in Manufacturing 18(5), 501–514 (2008)
6. Rachel, W.Y., Andy, C.L., Edwin, T.C.: The impact of employee satisfaction on quality and profitability in high-contact service industries. Journal of Operations Management 26, 651–668 (2008)
7. Nakade, K., Ohno, K.: An optimal worker allocation problem for a U-shaped production line. International Journal of Production Economics 60, 353–358 (1999)
8. Paquet, M., Martell, A., Montreuil, B.: A manufacturing network design model based on processor and worker capabilities. International Journal of Production Research 46(7), 2009–2030 (2008)
9. Valls, V., Perez, A., Quintanilla, S.: Skilled workforce scheduling in service centres. European Journal of Operations Research 193, 791–804 (2009)
10. Salman, A., Ahmad, I., Al-madani, S.: Particle swarm optimization for task assignment problem. Journal of Microprocessors and Microsystems 26, 363–371 (2002)
11. Sha, D.Y., Hsu, C.-Y.: A hybrid particle swarm optimization for job shop scheduling problem. Journal of Computers & Industrial Engineering 51, 791–808 (2006)
12. Yaakob, S.B., Watada, J.: Placement problem in an industrial environment: Knowledge-based intelligent information and engineering systems. In: Lovrek, I., Howlett, R.J., Jain, L.C. (eds.) KES 2008, Part III. LNCS (LNAI), vol. 5179, pp. 111–118. Springer, Heidelberg (2008)
13. Kennedy, J., Eberhart, R.C.: Particle swarm optimization. In: Proceedings of the IEEE International Conference on Neural Network, pp. 1942–1948 (1995)
14. Shi, Y., Eberhart, R.: Empirical study of particle swarm optimization. In: Proceeding of the Congress on Evolutionary Computation (CEC 1999), pp. 1945–1950 (1999)
15. Chen, W., Zhang, R.-T., Cai, Y.-M., Xu, F.-S.: Particle swarm optimization for constrained portfolio selection problems. In: Proceeding of the Fifth International Conference on Machine Learning and Cybernetics, pp. 2425–2429 (2006)

Evidential Reasoning Based on DNA Computation

Rohani Binti Abu Bakar and Junzo Watada

Graduate School of Information, Production and Systems, Waseda University
2-7, Hibikino, Wakamatsu-Ku, Kitakyushu-Shi, Fukuoka-Ken, 808-0135, Japan
rohani@ump.edu.my, junzow@osb.att.ne.jp
Tel.: +81-93-692-5179; Fax: +81-93-692-5179

Abstract. The objective of this study is to present an alternative approach to solve reasoning problems. DNA computing technique shows to solve evidential reasoning problems in this study. The reasoning is here executed on the basis of the concepts of plausibility and belief function. The evidential reasoning is a process dealing with problems that having both quantitative and qualitative criteria under various uncertainties including ignorance and randomness of information. The procedure to solve reasoning problem by means of DNA computing has been illustrated. An experiment shows the steadfastness of DNA computing to reach the solution of a reasoning problem.

Keywords: DNA computing, Reasoning, Evidential reasoning, Plausibility, Belief function.

1 Introduction

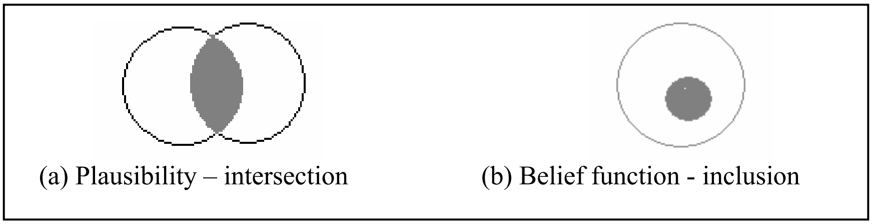
Reasoning is defined as the process of inferring conclusions from statements. The reasoning process has two main parts, *inference* and *conclusion*. *Inference* is the use of a rule or warrant to link some proposition (statements) with others meanwhile *conclusion* is the proposition toward which the inference moves. From the definition, reasoning can be the movement of evidence from propositions to propositions, but the conclusion of reasoning is an action.

From a normative point of view the study on the logic of reasoning is to judge whether reasoning is weak or strong, good or bad, valid, fallacious, etc. Reasoning normally has a direction, in a context of argument, toward some goal. But it does not always need to direction. There can be aimless reasoning. When reasoning is used to fulfill an affording proof or evidence, there is normally one particular proposition designated as the ultimate conclusion to be proved in the sequence of reasoning. This goal gives the sequence of reasoning its purpose.

1.1 Evidential Reasoning

The evidential reasoning approach is widely used in decision support system and expert system to help a decision maker in preparing possible solution for certain decision [6]. The theory of evidence is deals with weights of evidence and with the numerical degrees to support based on evidence [1].

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 212–219, 2009.



(a) Plausibility – intersection

(b) Belief function - inclusion

Fig. 1. Illustration of (a) plausibility and (b) belief function

1.2 Belief and Plausibility Concepts

Belief concept is referred to model the degree of someone's belief. Theory of belief functions is based on two ideas which are; (1) the idea of obtaining the degree of belief for one question from subjective probabilities for a related question, and (2) Demspter's rule for combining such degrees of belief when they are based on independent items of evidence [5][7].

Level of confidence is expressed as a percentage of instances that a set of similar constructed tests will capture the true mean of accuracy for the system being tested within a specified range of values around measured accuracy value of each test. Meanwhile priori probabilities expressed marginal probability to interpret a description of what is known about a variable in the absence of the same evidence [8][9]. It is largely employed in Bayesian statistical inferences. Figure 1 illustrates graphically the plausibility and belief function. Plausibility can be defined as the summation of all movable mass of intersected sets with the focal set A as shown in Figure 1(a), meanwhile belief function can be represent as the summation of all movable masses of sets included in the focal set A as shown in Figure 1(b).

The conventional technique to calculate plausibility and belief function requires complex steps and calculations. These complex steps and calculations directly require huge resources of either computing resources or human expertise when we consider a large number of data.

DNA computing is totally different, where all the calculation and process of considering feasible solutions are done in a single step during the process of ligation and hybridization [10]. Even though wet experiment of DNA requires quite huge amount of time even for processing a small size of data, but it is very effective if we consider a large number of data. This is because it spends the same processing time to deal with both small and large sizes of data in DNA computing. Almost all process in DNA computing can be solved by complexity of O_n meanwhile ligation and hybridization and gel electrophoresis can be done by O_1 , where O_n and O_1 denote the movement of the number of steps of executing DNA wet-experiment process.

2 DNA Computing Approach to Solve Reasoning Problems

2.1 DNA Procedures

In this section, we will briefly explain DNA procedure to solve reasoning problems. In solving reasoning problems through DNA procedure, we follows Demspter-Shafer's model when plausibility and belief of certain condition will be considered. As we discussed above, plausibility represents intersection relation of data meanwhile belief represents inclusion relation of data in universal set Ω [8]. Thus, we propose DNA procedure to find a plausibility and belief in one universe set. As the prerequisite of our proposed method, we assume that we have the knowledge of a target set to identify a plausibility and belief function for Ω .

As a prerequisite in DNA computing procedure, DNA should be synthesized to code data. In this study, we randomly generate ssDNA to represent each element in the set of Ω . The detail algorithm will be discussed further in the next sections.

2.2 An Example

In this section, we will provide an example to solve through DNA computing technique. The example is adopted from [2]. In this paper, the author has considered the breakable sensor. However, to make it more understandable, we simplified the breakable sensor problem. Details of explanation for the considered problem can be found in [2].

3 Model

In order to solve this problem, let us present it in a finite boolean algebra of propositions or of sets which here are equivalent as below

$$\Omega = S \times T \times \Theta \tag{1}$$

Where;

S = {B , R}, the sensor status, Blue or Red

T = {TH, TC}, the temperature status, Hot or Cold

Θ = {ThW, ThB}, the thermometer status, Working or Broken.

The eight elements of the space Ω are detailed as shown in Table 1.

Table 1. The labels of the eight elements in Ω

	B		R	
	TH	TC	TH	TC
ThW	a	b	c	d
ThB	e	f	G	h

Consider this example, degree of belief on subsets of Ω is quantified by a probability distribution P:2^{Ω} →[0,1] such that $\forall \omega \in \Omega$, $P(\{\omega\})$ and $\forall A$, $B \subseteq \Omega$ with $A \cap B = \phi$, $P(A \cup B) = P(A) + P(B)$ and $P(A) = \sum_{\omega \in A} p(\omega)$.

On the other hand, the probability distribution of P on 2^{Ω} for the above example can be calculated as below.

$$P(ThW) = P(\{a,b,c,d\}) = p(a) + p(b) + p(c) + p(d) = 0.8 \tag{2}$$

$$P(ThB) = P(\{e,f,g,h\}) = p(e) + p(f) + p(g) + p(h) = 0.2 \tag{3}$$

Equations (2) and (3) illustrate the information that the box induces the constraints for each possibility. Table 2 shows probability distribution on Ω based on events shown in Table 1.

Table 2. Probability distribution on Ω = $S \times T \times \theta$

	B		R	
	TH	TC	TH	TC
ThW	0	.8π	.8(1-π)	0
ThB	.2(1-π)	.2π x	.2(1-π)(1-x)	.2π (1-x)

Considered the transferable belief model analysis in [3], when π in unknown, the results in the basic belief masses can be shown as below:

m({a,b}) = 0.8 *and* m({e,f,g,h}) = 0.2

Conditioning on B, ThB implies the transfer of all basic belief masses within the set {a,b,e,f}. The basic belief masses m_B are:

m_B ({b}) = 0.8 *and* m_B({e,f}) = 0.2

Thus,

bel(TC|B) = bel_B ({b,f}) = 0.8 *and* pl(TC|B) = pl_B ({b,f}) = 1

On the other hand, when π in known, the results in the basic belief masses m' with

m' ({b}) = 0.8 π, m' ({c}) = 0.8 (1-π), m'({e,g}) = 0.2(1-π), m'({f,h}) = 0.2π

Conditioning on B implies the transfer of all basic belief masses within the set {a,b,e,f}. The updates basic belief masses m'_B are

m'_B ({b}) = 0.8 π m'_B ({e}) = 0.2(1-π) m'_B ({f}) = 0.2 π

Thus

$$bel' (TC|B) = bel'_B (\{b,f\}) = \frac{\pi}{0.8\pi + 0.2'}$$

and

$$pl'(TC|B) = pl'_B(\{b,f\}) = \frac{\pi}{0.8\pi + 0.2'}$$

4 Experiment

In this section, the experiment is designed for DNA computation to solve a reasoning problem. The above example is considered to explain how a DNA computing approach can solve reasoning problem. The concepts of believe and plausibility are engaged in this experiment to obtain the final result. The procedures in Section 2.1 are explained in this section to solve the above problem.

Step 1. At the first place, randomly 12-mer of ssDNA is generated to represent each value as shown in Table 1. The values that are represented by ssDNA are status of sensor {B, R}, temperature status {TH, TC} and thermometer status {ThW, ThB}. Table 2 shows the randomly generated 12-mer of ssDNA to represent all these values. For the purpose of generated ssDNA, the software of DNASequenceGenerator developed by Udo *et al.* [4] has been employed.

Step 2. From the following ssDNAs, the strands are synthesized to represent elements in Ω. For example, if a strand to represent that sensor status is Blue, temperature status is Hot and thermometer status is working, {B ∩ TH ∩ ThW} can represent the combination of second half 6-mer of B, 12-mer of TH and first half 6-mer of ThW. Figure 1 shows how this connection can be designed.

From the following ssDNAs, the set of Ω's elements are synthesized as follows: From (2) and (3), sets in space Ω (which represent sensor is working or broken) can be defined as below:

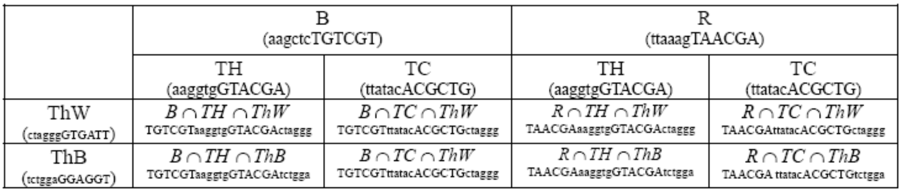
P(ThW) = P{(TGTCGTaaggtgGTACGActaggg),(TGTCGTttatacACGCTGctaggg),
(TAACGAaaggtgGTACGActaggg),(TAACGAttatacACGCTGctaggg)}

P(ThB) = P{(TGTCGTaaggtgGTACGAtctgga)(TGTCGTttatacACGCTGctaggg)
(TAACGAaaggtgGTACGAtctgga)(TAACGA ttatacACGCTGtctgga)}

Table 3. ssDNA of represented values in the case study

Values	ssDNA (3' → 5', 12 mer)
B	aagctctgtcgt
R	ttaaagtaacga
TH	aaggtggtacga
TC	ttatacacgctg
ThW	ctaggggtgatt
ThB	tctggaggaggt

Table 4. ssDNA fo rm of all possible elements in Ω to considered in this study



	B (aagctcTGTCGT)		R (ttaaagTAACGA)	
	TH (aaggtgGTACGA)	TC (ttatacACGCTG)	TH (aaggtgGTACGA)	TC (ttatacACGCTG)
ThW (ctagggGTGATT)	$B \cap TH \cap ThW$ TGTCGTaaggtgGTACGActaggg	$B \cap TC \cap ThW$ TGTCGTttatacACGCTGctaggg	$R \cap TH \cap ThW$ TAACGAaaggtgGTACGActaggg	$R \cap TC \cap ThW$ TAACGAttatacACGCTGctaggg
ThB (tctggaGGAGGT)	$B \cap TH \cap ThB$ TGTCGTaaggtgGTACGAtctgga	$B \cap TC \cap ThW$ TGTCGTttatacACGCTGctaggg	$R \cap TH \cap ThB$ TAACGAaaggtgGTACGAtctgga	$R \cap TC \cap ThB$ TAACGA ttatacACGCTGtctgga

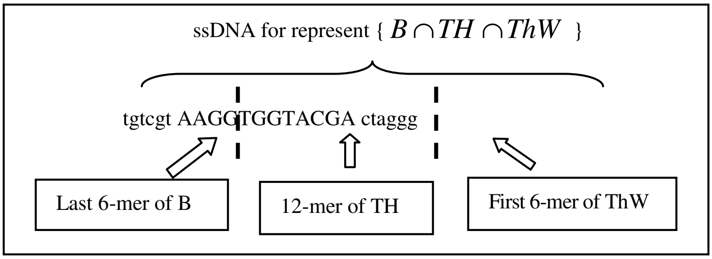

ssDNA for represent { $B \cap TH \cap ThW$ }
tgtcgt AAGGTGGTACGA ctaggg
Last 6-mer of B
12-mer of TH
First 6-mer of ThW

Fig. 1. Example of ssDNA synthesized to which are represent elements in Ω

From these two sets, set of Ω can be defined as follows:

Ω=
{(TGTCGTaaggtgGTACGActagggTGTCGTttatacACGCTGctagggTAACGAaaggtg
GTACGActagggTAACGAttatacACGCTGctaggg),

(TGTCGTaaggtgGTACGAtctggaTGTCGTttatacACGCTGctagggTAACGAaaggtgG
TACGAtctggaTAACGAttatacACGCTGtctgga)}

To solve reasoning problem, the plausibility and belief function should be obtained from Ω. In order to identify plausibility and belief function. As general, plausibility can be defined as the intersection between P (ThB) and P (ThW) in space Ω. Meanwhile, belief function can be defined as the inclusion of sets in space Ω.

In the next step, procedures to obtain plausibility and belief function are discussed. In the first part, the procedures to obtain the plausibility are discussed, and then procedures are explained to obtain the belief function.

Step 3 (plausibility)

i. Put each strand that represents each set from Ω in a different test tube, say strand P(ThW) is in test tube T_1 and labeled as S1' and strand P(ThB) is in test tube T_2 and labeled as S2'.

ii. Using affinity separation process, a complementary strand of specific ssDNA (e.g complementary of strand of TH, labeled as TH') is placed into magnetic beads. Pour the DNAs from T_1 onto magnetic beads.
   a. If there are ssDNA extracted, let us name it as E1 (E1 is an original ssDNA of TH that is extracted from T_1).
   b. If not any ssDNA extracted from T_1, change to other ssDNA.
   c. Repeat the process.

iii. To check the plausibility elements with T_2
   a. Put the complementary strand of E1, let us denote this complementary strand of E1 as E1' and stick it into a magnetic beads
   b. With affinity separation process, pour the solution from T_2 into magnetic beads.
   c. If there are ssDNA extracted, remove it to T_3.

iv. Repeat steps of (ii and iii) for all ssDNA from T_1 and T_2.

v. Read-out a left strands in T_1, T_2 and T_3.

vi. If all three test tube have a different strands, we can say that $T_1 \subseteq T_2$ and vice versa.

Thus, there are plausibility between element P(ThW) and P(ThB)

Step 4 (belief function)

i. Put all strands from Ω into a test tube, labeled as T_1.

ii. Synthesize a complementary strand for given strand (S) (say, we intend to check either $B \cap TC \cap ThW$ is belief function for Ω or not, therefore strand that represent $B \cap TC \cap ThW$ is labeled as S). Label the strand as S". Put strand of S" into magnetic beads for affinity separation process.

iii. Pour solution from T_1 onto magnetic beads of S".

iv. Extract the complement strand for S" from T_1.

v. Run the gel electrophoresis step to read the strand from S" and T_1, then readout.

vi. if S" < T_1, we can say S"∩ Ω.

5 Concluding Remarks

In this study, we have proposed a new approach to solve evidential reasoning with DNA computing technique. We have illustrated our proposed solution to tackle the simplified breakable sensor problem. In solving this problem with evidential reasoning approach, plausibility and belief function have been examined by DNA computing technique.

At the current stage, the proposed algorithm only able to examine the existences of belief functions and plausibility in the universal search space of Ω without calculate the degree of these two measurements. In the near future, we hope to upgrade the algorithm to include the calculation module in order to allow the calculation of belief functions and plausibility can be done in one process.

Solving reasoning problem by DNA computing technique can offer a new perspective of reasoning field where the huge size of data can be considered to solve and the super massive parallel.

Acknowledgments. The first author would like to thank University Malaysia Pahang and Kementerian Pengajian Tinggi Malaysia for supporting study leave.

References

[1] Shafer, G.: A Mathematical Theory of Evidence. Princeton University Press, Princeton (1976)

[2] Smets, Ph.: What is Dempster-Shafer's Model? In: Advance in The Dempster-Shafer Theory of Evidence, pp. 5–34. John Wiley & Sons, Inc., New York (1994)

[3] Smets, Ph., Kennes.: The Transferable Belief Model. Technical Report: TR- IRIDIA-90-14 (1994)

[4] Udo, F., Sam, S., Wolfgang, B., Hilmar, R.: DNA Sequence Generator: A Program for Constructing of DNA Sequences. In: Seventh International Workshop on DNA Based Computer, pp. 23–32 (2001)

[5] Walkins, E., Lavington, S.H.: Towards o Theory of Evidence, Internal Report CSM-325, Department of Computer Sciences, University of Essex, United Kingdom (1999)

[6] Mohamed, R., Watada, J.: Evidential Reasoning Evaluation of Hierarchical Structure. In: Proceeding of The 5th International Symposium on Management Engineering 2008, Kitakyushu, Japan, March 15-17, 2008, pp. 406–415 (2008)

[7] Shafer, G.: A Mathematical Theory of Evidence. Princeton University Press, New Jersey (1975)

[8] Yang, J.B., Wang, Y.M., Xu, D.L., Chin, K.S.: The Evidential Reasoning Approach for MADA under Both Probabilistic and Fuzzy Uncertainty. European Journal of Operational Research 171, 309–343 (2006)

[9] Saffiotti, A.: Issues of Knowledge Representation in Dempster-Shafer's Theory. In: Yager, R.R., Fedrizzi, M., Kacprsyk, J. (eds.) Advances in the Dempster-Shafer Theory of Evidence, pp. 415–440. Wiley, Chichester (1994)

[10] Adleman, L.M.: Molecular Computation of Solutions to Combinatorial Problems. Science 266(5187), 1021–1024 (1994)

Dynamic Tracking System through PSO and Parzen Particle Filter

Zalili Binti Musa, Junzo Watada, Sun Yan, and Haochen Ding

Graduate School of Information, Production and System
Waseda University
2-7 Hibikino, Wakamatsu, Kitakyushu, Fukuoka 808-0135 Japan
zalili@ump.edu.my, junzow@osb.att.ne.jp
an_pirlo21@hotmail.com, haochen.ding@gmail.com

Abstract. Transportation plays a pivotal role in our society, especially in a good quality of life and economic prosperity. Intelligent transportation system (ITS) has been developed to manage the transport infrastructure and vehicles since the number of vehicles is rapidly growing and to avoid any accident. Various applications have provided to support ITS. One of them is a driver-assistant system. Considering of heavy vehicles such as bus, truck, trailer and etc., the driver assistant system is of importance in monitoring and recognizing objects in vehicle surrounding. For example, in operating a heavy vehicle, a driver has a limited view of the vehicle surrounding itself. It is difficult for the driver to ensure that the surrounding of vehicle is safe before operating the machine. Thus, in this paper, we employ a video tracking system through PSO and Parzen particle filter to break through several problems such as simultaneous motion and occlusion among objects. This method makes it easy to track a human movement from every frame and indirectly require less a processing time for tracking an object location in a video stream compared to conventional method. The detail outcome and result are discussed using experiments of the method in this paper.

Keywords: Dynamic tracking system, PSO, template matching, Parzen particle filter, and human tracking.

1 Introduction

Transportation plays a pivotal role in our society, especially in realizing a good quality of life and economic prosperity. The economically and environmentally efficient transportation system is required to solve our transportation issues such as safety and security. Intelligent transportation system (ITS) has been developed to manage the transport infrastructure and vehicles since the number of vehicles are rapidly growing. This system employs various technologies such as information and communication technology including wireless communications, computational technologies, floating car data or floating cellular data, sensing technologies, inductive loop detection and video vehicle detection.

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 220–227, 2009.

Various applications have been provided to support ITS. One of them is a driver-assistant system. The objective of assistant systems is to detect obstacles surounding the vihecle and lanes on roads [1],[2],[3], monitor surrounding vehicles to avoid accidents [4],[5], find transportation speed limit signs to warn of over speeding [6],[7], and localize vehicles by landmarks [8].

Nowadays, driver assistant systems are utilized in many automobiles. This is because a driver safety has become an important and serious issue to prevent a vihecle from an accident. Considering of heavy vehicles such as bus, truck, trailer and etc., driver assistant system is important to monitor and recognize obstacles surrounding vehicles. For example, in operating the heavy vehicle, a driver has a limited view surrounding the vehicle. It is difficult for the driver to ensure that the surrounding of a vehicle is safe before operating the machine.

In this paper we employ a wide view camera installed on a large vehicle such as a bus to capture videos around the vehicle. The purpose of this research is to help a driver to avoid or mitigate an accident through detecting an object using multi-camera tracking system. A particle swarm optimization and Parzen filter algorithm based on template matching are proposed in order to detect and track a static and moving object that enables us to address the following problems:

1) ***Simultaneous motion***

   For this purpose, we installed the camera at the vehicle. Thus during the process of image capturing, the camera and object are moving concurrently. It is resulted in that the location and motion object are changing dynamically. Therefore, it is hard to measure a possibility object location during the tracking process.

2) ***Hidden objects or occlusion among objects***

   Hidden objects will occur during a camera tracks the objects from the location. Both of object motion and camera location actively give an influence for the occlusion.

In the remainder of this paper, we will describe in detail the techniques and algorithm that we use in our proposed method in Sections 2, 3 and 4. Our detailed outcomes and simulation results will be discussed in Section 5, and finally, we conclude with a discussion of the results and opportunities for future work in Section 6.

2 Target Detection

In target detection process, we combine a background subtraction and histogram method to detect the location of humans which at the first enters in the system. Background subtraction can be defined as the process of subtracting related information from the image and then finally, all information will be processed through threshold to obtain the object interested. Using $I \times J$ size image, the background subtraction is done by means of the following formula:

$$\mathrm{g}(x,y)=\begin{cases}1: \left|\mathrm{a}(x,y)-\mathrm{b}(x,y)\right| > \lambda \\ 0: \text{otherwise}\end{cases} \tag{1}$$

where $g(x, y)$ is a new pixel value, $a(x, y)$ a value of fixed background pixel and $b(x, y)$ a value of frame pixel. x and y are the coordinates of image pixel, where $x = 1,..,I$ and $y = 1,..,J$. λ is threshold value denoted as:

$$\lambda = \frac{\sum(a(x, y) - b(x, y)}{I \times J} . \tag{2}$$

A histogram method is applied to the image after the subtraction process. The purpose of this method is to recognize the location of objects based on x and y axes. In this method, we can divide the procedure into two sub-processes: (1) determining the location of x-axis (horizontal process) and (2) determining the location of y-axis (vertical process).

In the horizontal process, an image $g(x, y)$ will be divided into several smaller windows $l \times J$ size pixel. l is a integer number and $l < I$. After that, for each smaller window, $d(i, j)$ gets the sum of pixel value written in the following:

$$h(r) = \sum d(i, j) . \tag{3}$$

where $h(r)$ is the feature of a binary image, and i, j are the integer number of pixels in a smaller window. The peaks and their region of surrounding areas of $h(r)$ are extracted to produce a set of vertical slices of the image based on x-axis.

$$\partial x_{1:e} = \partial x_1 ,.., \partial x_e . \tag{4}$$

Next, the vertical process is executed from these slices to determine the object candidate regions based on y-axis. This process is the same process to determine the location of x-axis. Therefore, this process can produce a set of horizontal slices based on y-axis of the image.

$$\partial y_{1:e} = \partial y_1 ,.., \partial y_e . \tag{5}$$

Meanwhile, in our study we consider a center point of object tracking process. The coordinate of center point of object (x, y) can be defined as follows:

$$x = \frac{|\partial x_e - \partial x_{e+1}|}{2} \text{ and} \tag{6}$$

$$y = \frac{|\partial y_e - \partial y_{e+1}|}{2} . \tag{7}$$

3 Particle Swarm Optimization and Template Matching

During the process of an analysis, a target or visual object rapidly changes in any direction at each frame. Particle swarm optimization is employed with template matching to break through this problem.

After we found the region and center of coordinates of a human's location on the x and y axes using target detection process, the process of feature extraction is executed. At the first, we extract the feature of an object region by dividing into *n* smaller windows. For each small window, we employed the equation (8) to calculate the feature space, F as shown below:

$$F_i = \frac{1}{2\sigma}\exp\left(-\frac{|vx-\mu|}{\sigma}\right) \tag{8}$$

where σ and μ are the mean values of the whole region and a small window, respectively. vx is a gray level value of the object region and $i=[1,...,n]$ are the integer numbers of small windows. Suppose that we have several object regions in feature extraction process and denoted as $p(j)=[F_1,...,F_n]$ where j is the numbers of objects.

In this research, the PSO algorithm is applied to find out the minimum values of template matching process. Assuming an image has an $M \times N$ size, we initialize the position of m particles as formula below:

$$f(l)=[px_l, py_l] \tag{9}$$

where $l=[1,...,m]$ is a particle, px and py are a random locations based on image size and denoted as:

$$px = rand(1,m)\times M \tag{10}$$

$$py = rand(1,m)\times N \tag{11}$$

In order to achieve a minimize value, we denote the number of iteration by η. According to this, the best position, $pbest$ is obtained based on the following rule:

$$pbest_\eta(l)=\begin{cases} f_\eta(l):(\tau_\eta(l)<\tau_\eta(l-1)) \\ pbest_\eta(l):(\tau_\eta(l)>\tau_\eta(l-1)) \end{cases} \tag{12}$$

where τ is Euclidean distance value that we get from template matching process. In order to measure the matching process, we applied the Euclidean distance as shown in the following equation:

$$\tau_\eta(l)=\sqrt{\sum(p(j)-q_\eta(l))^2} \tag{13}$$

where $q_\eta(l)$ is a new feature from the image. On the other hand, in this case of $\tau_\eta(l)>\tau_\eta(l-1)$, current value τ can be defined as below:

$$\tau_\eta(l)=\tau_\eta(l-1) \quad if(\tau_\eta(l)>\tau_\eta(l-1)) \tag{14}$$

Furthermore, the global best position, $gbest$ can be defined as below:

$$gbest_\eta(l)=pbest_\eta(l) \quad if\ l == location(\min(\tau_\eta)) \tag{15}$$

The current position, $f(l)$ and velocity, v are updated after each iteration using

$$v(l+1)=\omega v(l)+c_1r_1(pbest(l)-f(l))+c_2r_2(gbest(l)-f(l)) \tag{16}$$

and

$$f(l+1)=f(l)+v(l+1) \tag{17}$$

where ω is a inertia weight, c_1 and c_2 are constant values, r_1 and r_2 are random values, respectively. In our research, we initialize the values 0.2 and 1.4 to inertia weight and constant values, respectively.

4 Parzen Particle Filter

In general, the parzen particle filter is executed based on Kernel density estimator. Using this density estimator, the performance of the traditional algorithm in broader kernel can be improved [9].

In our study, the state space of each target includes both position axes *x* and *y* denoted as:

$$\mathrm{X}_t^k = \left[x_t^k, y_t^k\right] \tag{18}$$

where, x and y are coordinates of a pixel image. At the first step (time=0), a set of particles, S_t^k is initialized to 0:

$$\mathrm{S}_t^k = [\mathrm{X}_t^k \mathrm{W}_t^k : k = 1 \ldots N] = 0 \tag{19}$$

In our case, an observation process is corresponding to observe a new position of object movement for every target. The observation Y_t^k is obtained by:

$$\mathrm{Y}_t^k = \left[x_t^k, y_t^k\right] \tag{20}$$

Based on the equation above, the joint distribution of X and Y is the distribution of the intersection of the events X and Y, that is, the one of the co-occuring event of X and Y. However, the main concern is to find the conditional density of the state X_t at time t, which can be represented by $P(\mathrm{X}_t | \mathrm{Y}_{1:t})$. This conditional density can be obtained by predicting and updating recursively.

In our study, we used a Parzen particle filter. In this method each kernel can be propagated through mapping $P(\mathrm{X}_t | \mathrm{X}_{t-1}^k)$ by using a local linearization and can be addressed as continuous output distribution $P(\mathrm{X}_t | \mathrm{Y}_{1:t})$. $P(\mathrm{X}_t | \mathrm{Y}_{1:t})$ can be generated by the process of noise sampling as written in the following:

$$P(\mathrm{X}_t | \mathrm{Y}_{1:t}) = \sum_t^N \mathrm{W}_t^k \mathrm{K}(\mathrm{A}_t^k (\mathrm{X}_t - \mathrm{X}_t^k)) \tag{21}$$

where K is taken a standard Gaussian function with mean zero and variance 1 and A is a transformation matrix. Transformation matrix A can be calculated as:

$$\mathrm{A}_t^k = \mathrm{A}_{t-1}^k \mathrm{J}\Big|_{\mathrm{X}_{t-1}^k}^{-1} \tag{22}$$

Meanwhile, J is defined by Jacobian:

$$\mathrm{J}\Big|_{\mathrm{X}_{t-1}^k} = \frac{\partial \mathrm{f}}{\partial \mathrm{x}}\Big|_{\mathrm{X}_{t-1}^k} \tag{23}$$

Furthermore, to update a weight sample two processes are involved; weight evaluation and weight normalization. In weight evaluation, the weight sample can be drawn by:

$$\mathrm{W}_t^k = \mathrm{W}_{t-1}^k P(\mathrm{Y}_t|\mathrm{X}_t^k)\left|\mathrm{J}\right|_{\mathbf{X}_{t-1}^k}\Big| \tag{24}$$

For the likelihood we used a function related to noise sample, where we assumed a Gaussian process with zero mean and a variance. Therefore, we can denote as:

$$P(\mathrm{Y}_t|\mathrm{X}_t^k) = \mathrm{W}_{t-1}^k \frac{\exp\left(\frac{1}{\alpha}\left(\mathrm{Y}_{t-1} - \mathrm{X}_{t-1}^k\right)\sum\left(\mathrm{Y}_{t-1} - \mathrm{X}_{t-1}^k\right)\right)}{\sqrt{2\pi\alpha}} \tag{25}$$

In the other hand, for $k = 1,...,N$, normalize the importance weights as in the following formula:

$$\bar{\mathrm{W}}_t^k = \mathrm{W}_t^k \left[\sum_{j=1}^{N} \mathrm{W}_t^{(j)}\right]^{-1} \tag{26}$$

At the final stage, the output of the particle filter is given by:

$$\bar{\mathrm{x}}_t = \sum_{k=1}^{S} \mathrm{W}_t^k \cdot \mathrm{X}_t^k \tag{27}$$

However, for re-sampling the detail explanation of this formula can be found in [10]. After re-sampling process, basically all the processes are repeated.

5 Result and Discussion

In order to evaluate the proposed method, we employed a wide FOV camera installed on a movement or stopping vehicle to capture videos around the vehicle. In this case, the images are captured by a fisheye camera. Two different types of video data have been tested (i) single human movement and (ii) plural human movements with different direction. One video stream was taken about 5 seconds (150 frames), and the size of the video frame is 320 x 240 pixels as shown in Fig.1.

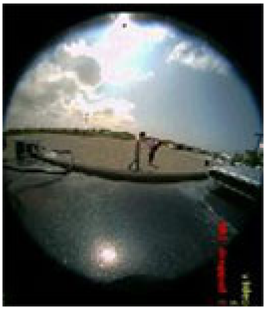


Fig. 1. Example of video image

For detection process, the result shows that our proposed method can recognize the human location well as shown in Fig. 2. Furthermore, the result shows the target is clearer and it is easy to recognize the location.

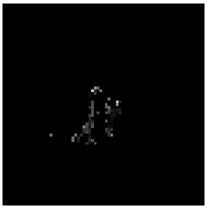


Fig. 2. Detection results

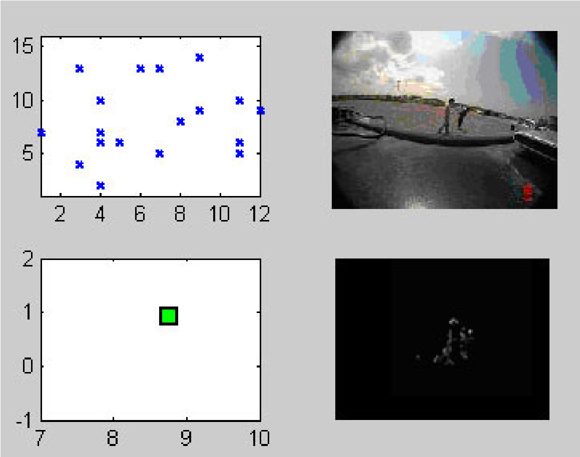


Fig. 3. Tracking result based on our proposed method

Besides that, using the PSO algorithm and Parzen particle filter is capable of identifying the human location and tracking several human motions for every frame as shown in Fig. 3. In addition, our proposed method is capable of tracking human movements during or after the overlapping between objects occurs compared to continuous detection method. It is able to discriminate two different object movements even when both of these objects are overlapped.

6 Conclusions

There are two main issues for human tracking system in a large view case. First is location detection and second is to track a human movement in every frame. In this paper, background subtraction and histogram methods are implemented in detection module to detect a human location. Meanwhile for tracking module, we proposed a new method using PSO and Parzen particle filter to solve two listed problems which we discussed in Section 1.

In this study, several video data such as single and multiple human movements with different direction have been tested. Furthermore, background subtraction is able to abstract features of image clearer and at this stage the features of the images are more prominent to our naked eyes. Therefore, it is easier for histogram method to recognize human locations precisely. In the other hand, our tracking module in the proposed method is capable to track the location of human in very stable and deals with different direction movement as we presented the result as written above.

Acknowledgments. The first author would like to thank to University Malaysia Pahang and Kementerian Pengajian Tinggi Malaysia for a study leave.

References

1. Broggi, A., Bertozzi, M., Fascioli, A., Guarino, C., Piazzi, A.: Visual perception of obstacles and vehicles for platooning. IEEE Trans. Intell. Transp. Syst., 164–176 (2000)
2. Chapuis, R., Aufrere, R., Chausse, F.: Accurate road following and reconstruction by computer vision. IEEE Trans. Intell. Transp. Syst., 261–270 (2002)
3. Kim, Y.U., Oh, S.Y.: Three-feature based automatic lane detection algorithm (TFALDA) for autonomous driving. IEEE Trans. Intell. Transp. Syst., 219–225 (2003)
4. Sun, Z., Bebis, G., Miller, R.: On-road vehicle detection using evolutionary Gabor filter optimization. IEEE Trans. Intell. Transp. Syst., 125–137 (2005)
5. Kato, T., Ninomiya, Y., Masaki, I.: Preceding vehicle recognition based on learning from sample images. IEEE Trans. Intell. Transp. Syst., 57–68 (2002)
6. Escalera, A., Armingol, J.M., Pastor, J.M., Rodriguez, F.J.: Visual sign information extraction and identification by deformable models for intelligent vehicles. IEEE Trans. Intell. Transp. Syst., 57–68 (2004)
7. Wu, W., Chen, X., Yang, J.: Detection of text on road signs from video. IEEE Trans. Intell. Transp. Syst., 378–390 (2005)
8. Li, S., Hayashi, A.: Navigation by integrating iconic and GPS information. In: Proc. IEEE Int. Conf. Intell. Veh., pp. 213–218 (1998)
9. Ye, Z., Liu, Z.-Q.: Tracking Human Hand Motion Using Genetic Particle Filter. In: IEEE International Conference on Systems, Man and Cybernetics (SMC 2006), pp. 4942–4947 (2006)
10. Arulampalam, M.S., Maskell, S., Gordon, N., Clapp, T.: A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. IEEE Transactions on Signal Processing (IEEE Transactions on Acoustics, Speech, and Signal Processing), 174–188 (2002)

Text Mining for Customer Enquiries in Telecommunication Services

Motoi Iwashita[1], Shinsuke Shimogawa[1], and Ken Nishimatsu[2]

[1] NTT Service Integration Laboratories,
Midori-cho 3-9-11 Musashino, Tokyo, 180-8585, Japan
{iwashita.motoi,shimogawa.shinsuke}@lab.ntt.co.jp
[2] NTT East, Network Business Headquarters,
Nishi-Shinjuku 3-19-2 Shinjuku, Tokyo, 163-8019, Japan
k.nishimatsu@east.ntt.co.jp

Abstract. Analyzing failure trends and establishing effective coping processes for complex problems in advance is essential in telecommunication services. We propose a method for semantically analyzing and classifying customer enquiries efficiently and precisely. Our method can also construct semantic content efficiently by extracting related terms through analysis and classification. This method is based on a dependency parsing and co-occurrence technique to enable classification of a large amount of unstructured data into patterns because customer enquiries are generally stored as unstructured textual data.

Keywords: Semantic content, Co-occurrence, Dependency parsing, Telecom operation, Text mining.

1 Introduction

The increasing use of Fiber-To-The-Home (FTTH) and the Asymmetric Digital Subscriber Line (ADSL) have induced the expansion of a variety of services, such as the exponential increase in use of the Internet, the provision of Voice over Internet Protocol (VoIP), and video distribution services. Therefore, the end-to-end network structure has become complicated compared with a network providing a conventional fixed telephone service, if we consider the connection of home equipment, such as a modem, and the setup of its software. As a result, customer enquiries about problems, such as no-connection to an Internet/phone service, will increase and discovering the causes of problems becomes difficult.

Customer satisfaction decreases when a long time is spent on repairing the problem because discovering the cause is difficult. Therefore, it is important to analyze the failure trends and establish effective coping processes for complex problems in advance of receiving the customer enquiries. A text mining technique, such as morphological analysis, syntax analysis, or co-occurrence relation, is effective [1],[2],[3],[4] because the customer enquiry is unstructured data. This technique is applicable to customer questionnaire analyses in product development, word searches in portal sites, term frequency analyses in web

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 228–235, 2009.

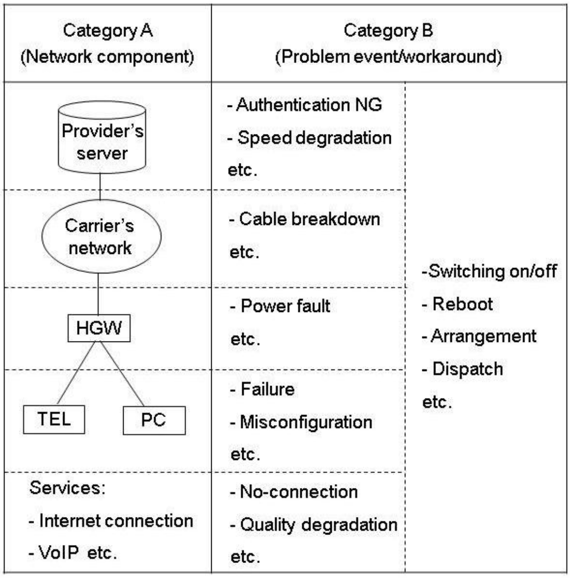
logs (blogs) or customer-generated media, article classification by keyword in news articles, and evaluation indexes of a company's image. Mainly morphological analysis is applied in these areas to survey trends by analyzing the frequency of terms in selected text. Clustering methods based on supervised learning have also been proposed [5],[6],[7],[8]. They are mainly applied for searching research papers by abstract and for automated scoring of descriptive answers, and they are effective for searching for similarities in text based on the given trend terms/information. Text similarity evaluation methods are introduced in Refs. [9], [10], [11], [12] in terms of sequential patterns of sentences, or the term frequency of the same word. These approaches are valid for verifying the similarity of classified texts, but knowing the semantic content of the text is impossible. Simply understanding failure trends and noting customer requirements when analyzing an enquiry then analyzing the word trends is not always effective. Because a telecom operator writes down information about a customer enquiry, the style of the description deeply depends on the operator. That is, the textual data contains of several sentences including failure points, failure phenomena, cause of the trouble, or a workaround process, such as "PC trouble induces no-connection", "blinking affects Internet trouble", and "switching on and off power supply because of modem blinking". Therefore, accurately understanding the meaning of sentences is essential. There are no effective and practical methods to semantically analyze text that are applicable to telecom management.

A text classification method that considers the features of telecom services and the co-occurrence of terms with dependency parsing for classifying and analyzing a large amount of unstructured data consisting of customer enquiries has been proposed [13]. We improve this method by adding detailed classification by telecom features considering not only the failed network component and problem event but also the affected network component and workaround. We can essentially construct the semantic content through texts classified by this function. Therefore, it can support the efficient establishment of coping processes. The features of textual data in telecom services are explained in Sec. 2. Our proposed classification and construction method is in Sec. 3. The evaluation results are described in Sec. 4.

2 Features of Textual Data in Telecom Services

Telecommunication service is generally provided by an end-to-end network consisting of a telephone, PC, the carrier's network, the provider's server, etc. as shown in Fig. 1.

There is clearly an event feature for each element of the network. We can predict that the equipment, such as the telephone, PC, or network, is strongly related to the problem, such as failure, misconfiguration of set-up, and cable breakdown, respectively. Moreover, an operator recommends to the customer an efficient workaround, such as switching the power supply on and off, rebooting the PC, or dispatching the problem to the provider/vendor. Therefore,



Category A (Network component)	Category B (Problem event/workaround)	
Provider's server	- Authentication NG - Speed degradation etc.	-Switching on/off - Reboot - Arrangement - Dispatch etc.
Carrier's network	- Cable breakdown etc.	
HGW	- Power fault etc.	
TEL PC	- Failure - Misconfiguration etc.	
Services: - Internet connection - VoIP etc.	- No-connection - Quality degradation etc.	

Fig. 1. Telecom features categorization

by designating the network factor as one event (Category A) and the problem/workaround as the other event (Category B), we can construct a semantic representation.

Next, we discuss the textual structure of a customer enquiry. To understand the customer enquiry and construct semantic content, including four kinds of information is desirable, such as troubled equipment, problem event, affected equipment/service, and workaround/related problem event, as shown in Fig. 2.

How much information is input deeply depends on the operators themselves. There are six patterns for a two-term relationship when we select two terms. Six examples, "PC is misconfigured", "modem trouble induces Internet trouble", "PC trouble induces no-connection", "blinking affects Internet trouble", "switching on and off the power supply because of blinking", "Internet is not connected", correspond to P21, P22, P23, P24, P25, and P26, respectively. We can construct the meaning of the classified text, but we need some sentences to fill areas not covered by the two terms.

We have four patterns for a three-term relationship when we select three terms. Four examples, "PC misconfiguration causes Internet trouble", "switching on and off the power supply because of the modem blinking", "PC trouble induces no-connection to the Internet", and "reinstalling software to the PC because of security trouble" correspond to P31, P32, P33, and P34, respectively. It is possible to classify the texts more precisely using three terms. These patterns can support the construction of semantic content among classified texts and establishment of coping processes.

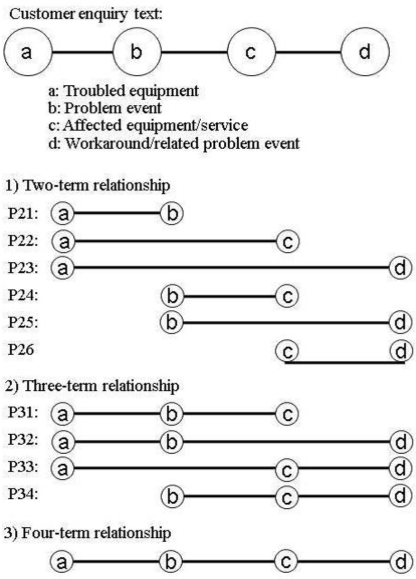


Fig. 2. Textual structure for telecom enquiry

3 Classification Method

The requirement for the classification of customer enquiries and construction of semantic content is based on the ability to cover all textual data and to match the operator's meaning. We want this classification to be determined from the viewpoints of term frequency, co-occurrence, and cause-effect relationship. Term frequency can tell us what kind of customer enquiries often appear, while co-occurrence tells us which terms strongly correlate. The cause-effect relationship tells us the relationship among multiple terms. We also want this construction of semantic content to be determined by the comparison of the co-occurrence rate among categorized and selected terms. The framework in which customer enquiries (textual data) are classified and the meaning in terms of the previous three criteria is constructed is shown in Fig. 3.

A large number of customer enquiries are input. Morphological analysis and dependency parsing is an effective preprocessing step for classification. Then, the terms are classified in terms of category as an input condition and calculated by frequency. Co-occurrences of these selected terms are calculated and selected by the threshold parameter as an input. Then we select the relationship between multiple terms by transition-rate calculation. Finally, we construct the semantic content for the classified texts to establish coping processes.

The details of the method from morphological analysis to transition-rate calculation are described in Ref. [13]. The semantic content is constructed through comparing the co-occurrence rate among terms. The three-term relationship is described below as an example.

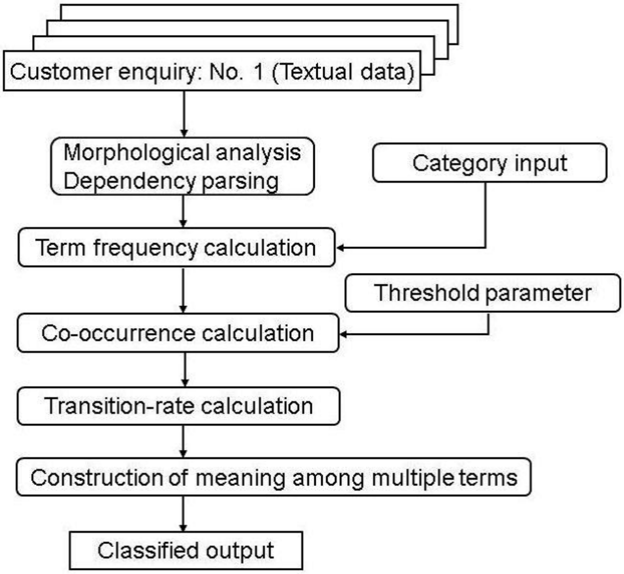

Customer enquiry: No. 1 (Textual data)
Morphological analysis
Dependency parsing
Category input
Term frequency calculation
Threshold parameter
Co-occurrence calculation
Transition-rate calculation
Construction of meaning among multiple terms
Classified output

Fig. 3. Framework of classification

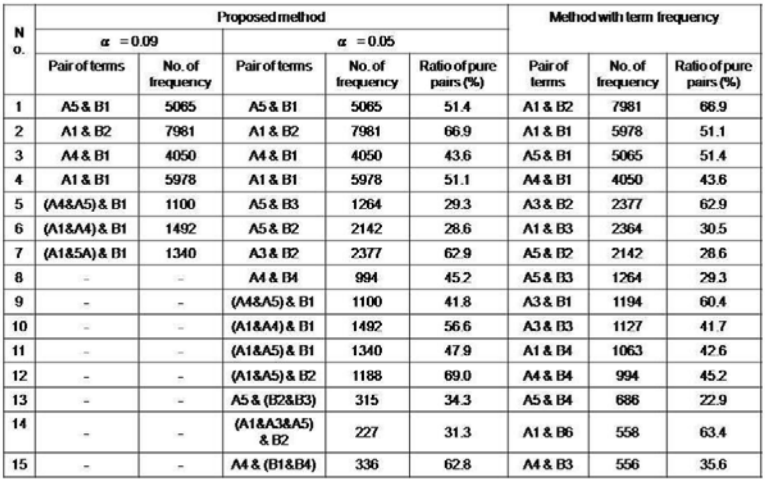
- Procedure 1: Classification of text into Pattern 1 (P31, P33) or 2 (P32, P34)
  - If the selected terms are, A_i and A_j, in category A (network component) and B_k, in category B (problem event/workaround), then go to Procedure 2, else Procedure 3.
- Procedure 2: Semantic content construction for Pattern P31 or P33
  - If $\alpha_1 > \alpha_2$, then $(A_i \& B_k) \rightarrow A_j$, or $A_j \rightarrow (A_i \& B_k)$
  - If $\alpha_1 < \alpha_2$, then $(A_j \& B_k) \rightarrow A_i$, or $A_i \rightarrow (A_j \& B_k)$
  α_1 and α_2 are the co-occurrence rates of $A_i \& B_k$ and $A_j \& B_k$, respectively.
- Procedure 3: Semantic content construction for Pattern P32 or P34
  - If $\alpha_1 > \alpha_2$, then $(B_i \& A_k) \rightarrow B_j$, or $B_j \rightarrow (B_i \& A_k)$
  - If $\alpha_1 < \alpha_2$, then $(B_j \& A_k) \rightarrow B_i$, or $B_i \rightarrow (B_j \& A_k)$
  α_1 and α_2 are the co-occurrence rates of $B_i \& A_k$ and $B_j \& A_k$, respectively.

The patterns described in Sec. 2 are selected in each classification.

4 Evaluation of Results

We used about 180,000 customer enquiries made during a month in 2007. We confirmed that the results obtained using 180,000 data items were almost the same as the results obtained using 18,000 data items from three days. Therefore, one month of data was stable. Co-occurrence among categories was calculated for the textual data from about 180,000 enquiries under the given thresholds $\alpha = 0.05, 0.09$. Pairs of terms obtained by our proposed method are shown in Table 1 with the pairs of terms obtained by term frequency. The number of frequency means how many times a pair of terms appeared in 180,000 enquiries.

Table 1. Comparison of two methods



N o.	Proposed method					Method with term frequency		
	α =0.09		α =0.05					
	Pair of terms	No. of frequency	Pair of terms	No. of frequency	Ratio of pure pairs (%)	Pair of terms	No. of frequency	Ratio of pure pairs (%)
1	A5 & B1	5065	A5 & B1	5065	51.4	A1 & B2	7981	66.9
2	A1 & B2	7981	A1 & B2	7981	66.9	A1 & B1	5978	51.1
3	A4 & B1	4050	A4 & B1	4050	43.6	A5 & B1	5065	51.4
4	A1 & B1	5978	A1 & B1	5978	51.1	A4 & B1	4050	43.6
5	(A4&A5) & B1	1100	A5 & B3	1264	29.3	A3 & B2	2377	62.9
6	(A1&A4) & B1	1492	A5 & B2	2142	28.6	A1 & B3	2364	30.5
7	(A1&5A) & B1	1340	A3 & B2	2377	62.9	A5 & B2	2142	28.6
8	-	-	A4 & B4	994	45.2	A5 & B3	1264	29.3
9	-	-	(A4&A5) & B1	1100	41.8	A3 & B1	1194	60.4
10	-	-	(A1&A4) & B1	1492	56.6	A3 & B3	1127	41.7
11	-	-	(A1&A5) & B1	1340	47.9	A1 & B4	1063	42.6
12	-	-	(A1&A5) & B2	1188	69.0	A4 & B4	994	45.2
13	-	-	A5 & (B2&B3)	315	34.3	A5 & B4	686	22.9
14	-	-	(A1&A3&A5) & B2	227	31.3	A1 & B6	558	63.4
15	-	-	A4 & (B1&B4)	336	62.8	A4 & B3	556	35.6

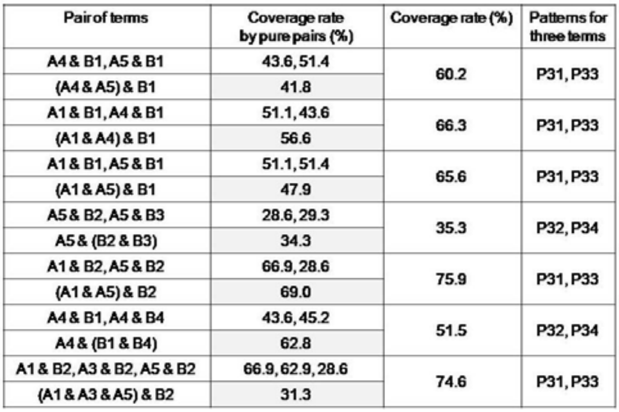
The ratio of pure pairs means the number of enquiries including only selected terms.

Let us first focus on the choice of pairs with the threshold as a parameter. The number of candidates increases when the threshold decreases. This is because of the weakness of co-occurrence. The number of pairs with a transition rate grows when the threshold decreases. Terms A1, A2, A3, A4 and A5 correspond to "internet", "e-mail", "IP-Tel", "PC" and "Modem" respectively, while B1, B2, B3 and B4 correspond to "misconfiguration", "blinking", "power supply on and off", and "security trouble".

Let us compare the proposed method and the method using only term frequency for selecting pairs of terms. As for the threshold value, the frequency of the 5th choice was 2,377 obtained by pair of terms frequency, while that of the 5th choice obtained by the proposed method was 1,100 with $\alpha = 0.09$. The choice obtained by term frequency was appropriate in this case because the difference between these two choices was large. Therefore, α should be improved to small value. The frequency of the 9th choice was 1,194 by pair of terms frequency, while that by the proposed method is 1,100 when $\alpha = 0.05$. However, the difference was small when $\alpha = 0.05$. Therefore, the choice obtained by the co-occurrence is appropriate and α can be determined by the difference between frequency by term frequency and that by the proposed method.

The selected pairs of two terms by the proposed method had a high co-occurrence rate and term frequency. We compared the pairs of three terms obtained by the proposed method and the pairs of two terms obtained using term frequency from the 9th choice to 15th choice. Using pair of two terms can not always construct applicable semantic content, this is because number of frequency by two terms is sometimes small compared with that by three terms (A3 & B3, and (A1 & A4) & B1, for example as shown in Table 1). Moreover, the coverage rate by pure pairs, which means text consisting of only selected terms,

Table 2. Coverage rate with co-occurrence



Pair of terms	Coverage rate by pure pairs (%)	Coverage rate (%)	Patterns for three terms
A4 & B1, A5 & B1	43.6, 51.4	60.2	P31, P33
(A4 & A5) & B1	41.8		
A1 & B1, A4 & B1	51.1, 43.6	66.3	P31, P33
(A1 & A4) & B1	56.6		
A1 & B1, A5 & B1	51.1, 51.4	65.6	P31, P33
(A1 & A5) & B1	47.9		
A5 & B2, A5 & B3	28.6, 29.3	35.3	P32, P34
A5 & (B2 & B3)	34.3		
A1 & B2, A5 & B2	66.9, 28.6	75.9	P31, P33
(A1 & A5) & B2	69.0		
A4 & B1, A4 & B4	43.6, 45.2	51.5	P32, P34
A4 & (B1 & B4)	62.8		
A1 & B2, A3 & B2, A5 & B2	66.9, 62.9, 28.6	74.6	P31, P33
(A1 & A3 & A5) & B2	31.3		

was evaluated as shown in Table 2. Using pairs of two terms for example, the coverage rates of A4 & B1 and A5 & B1 are 43.6% and 51.4 % respectively. If we consider (A4 & A5) & B1, the coverage rate becomes 60.2% by A4 & B1, A5 & B1, and (A4 & A5) & B1. The coverage rate mostly improved more than 50% when using pairs of three terms, while the pairs of two terms by term frequency did not reach a high coverage rate. The constructed semantic content is sometimes ambiguous by the pairs of two terms, while the pairs of three terms can construct semantic content more concretely. Therefore, introducing pair of three terms by the proposed method is effective for classifying semantic content.

Finally, we discuss the validity of constructing semantic content by the selected three terms. We calculated the concordance rate of the constructed meaning and the classified texts by applying the method explained in Sec. 3. In the case of (A4 & A5) & B1 corresponding to patterns P31 or P33, a high concordance rate of 81.5% was obtained. We also got a high concordance rate for the other six pairs. This suggests that the coping processes can be efficiently established.

5 Conclusion

A classification technique for customer enquiries is needed due to the increasing complexity of the connections in end-to-end networks in the telecom operating field. We proposed a method for analyzing and classifying customer enquiries that enables quick and efficient responses. Because customer enquiries are generally stored as unstructured textual data, this method is based upon dependency parsing and co-occurrence techniques to enable classification of a large amount of unstructured data into patterns. Moreover, a construction method of semantic content based on a comparison of the co-occurrence rate among selected terms was proposed. We applied the proposed method to 180,000 customer enquiries and evaluated its effectiveness.

We are currently conducting further studies on applying this method to real telecom operation in an operation support system.

References

1. Ohsumi, N.: Mining of textual data. Recent trend and its direction, http://wordminer.comquest.co.jp/wmtips/pdf/20060910_1.pdf
2. Sato, S., Fukuda, K., Sugawara, S., Kurihara, S.: On the relationship between word bursts in document streams and clusters in lexical co-occurrence networks. IPSJ 48-SIG14, 69–81 (2007)
3. Sullivan, D.: Document Warehousing and Text Mining. John Wiley, Chichester (2001)
4. Toda, H., Kataoka, R., Kitagawa, H.: Clustering news articles using named entities. IPSJ SIG Technical Report, 2005-DBS-137, pp.175–181 (2005)
5. Takahashi, S., Takahashi, S., Yasuda, N., Takahata, N., Ishikawa, T.: A Meaningful Keywords Extracting system based on A Sentence-Semantic Analysis Method. In: IPSJ, AI TR, vol. 90-8, pp. 65–72 (1992)
6. Akiba, Y., Tanaka, T., Suyama, T., Nagata, M.: Grading Examninee's Answer Sentences by Verifying Syntactic and Semantic Compatibility. In: IPSJ, SIG TR, 2006-NL-174(b), pp. 31–35 (2006)
7. Burnstein, J., Kukich, K., Wolff, S., Lu, C., Chodorow, M., Braden-Harder, L., Harris, M.D.: Automated scoring using a hybrid feature identification technique. In: Proc. of Thirty-Sixth Annual Meeting of the Association for Computational Linguistics and Seventeenth International Conference on Computational Linguistics, ACL-COLING 1998, pp. 206–210 (1998)
8. Taira, H., Mukouchi, T., Haruno, M.: Text Categorization Using Support Vector Machine. In: IPSJ, NL TR, 128-24, pp.173–180 (1998)
9. Sato, I., Nakagawa, H.: Mining Semi-structure for Text with Dependency Structure. In: IPSJ, SIG TR, 2006-DBS-140(II), pp. 207–214 (2006)
10. Agrawal, R., Srikaut, R.: Mining Sequential Patterns. In: Proc. of ICDE 1995, pp. 3–14. IEEE Computer Society Press, Los Alamitos (1995)
11. Manning, C.D., Schutze, H.: Foundations of Statistical Natural Language Processing. The MIT Press, Cambridge (1999)
12. Kawatani, T.: Document Clustering via Commonality Analysis of Multiple Documents. In: IPSJ, NL TR, 154-14, pp. 93–100 (2003)
13. Iwashita, M., Nishimatsu, K., Shimogawa, S.: Semantic analysis method for unstructured data in telecom services. In: Proc. of 2008 IEEE International Conference on Data Mining Workshops, pp. 789–795 (2008)

Defuzzification Using Area Method on L^∞ Space*

Takashi Mitsuishi[1] and Yasunari Shidama[2]

[1] University of Marketing and Distribution Sciences, Kobe 651-2188, Japan
Takashi_MITSUISHI@red.umds.ac.jp
[2] Shinshu University, Nagano 380-8553, Japan

Abstract. The mathematical framework for studying of a fuzzy approximate reasoning is presented. One of the defuzzification methods besides the center of gravity method which is the best well known defuzzification method is described. The continuity of the defuzzification methods and its application to a fuzzy feedback control are discussed.

Keywords: Approximate reasoning, Fuzzy control, Optimization, Functional analysis.

1 Introduction

Since 1990's, like the theory of classical control and modern control, many systematized mathematical considerations have been discussed [1]–[4]. In practical use, fuzzy membership functions (fuzzy sets), which represent input and output states in optimal control system, are decided on the basis of the experience of experts before. Therefore some acquisition methods of fuzzy inference rules by a neural network and a genetic algorithm have been proposed [5] [6].

The authors consider fuzzy optimal control problems as problems of finding the minimum (maximum) value of the cost (benefit) function with feedback law constructed by Mamdani method, product-sum-gravity method, and Nakamori method [7] [8]. These approximate reasoning methods adopt the center of gravity method, and calculate defuzzified value of inference result represented by fuzzy set. This defuzzification method is most widely used. The resulting behavior of fuzzy approximate reasoning using any of these defuzzification methods will be discussed in the following section. The author's study covers the area defuzzification method [9] [10]. Since this method does not synthesize the fuzzy set (membership function), this is better than the center of gravity method in respect of high-speed computing.

In this study, two kinds of continuity of defuzzification are discussed. One is Lipschitz continuity on the space of premise valuable. The other is continuity as functional on the set of membership functions. By the continuity as functional and the compactness of the set of membership functions in L^∞ space, the

* The paper was supported in part by Grant-in-Aid for Young Scientists (B) #19700225 from Japan Society for the Promotion of Science (JSPS).

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 236–243, 2009.

existence of an optimal feedback control law in a nonlinear fuzzy feedback control system, in which the feedback laws are determined by IF-THEN type fuzzy rules, are shown. Then it is crucial to investigate the convergence of feedback laws constructed by fuzzy approximate reasoning method and the convergence of solutions of the nonlinear state equation in the fuzzy control system.

Throughout this paper, $\mathbb{R}^n$ denotes the n-dimensional Euclidean space with the usual Euclidean norm $\|x\| = \left(\sum_{i=1}^{n} |x_i|^2 \right)^{\frac{1}{2}}$, $x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$.

2 Fuzzy Logic Control

2.1 IF-THEN Type Fuzzy Control Rules and Fuzzy Controller

The following IF-THEN type fuzzy control rules are considered in this study.

$$\text{Rule } i\text{: } IF\ x_1\ is\ A_{i1}\ and\ \ldots\ and\ x_n\ is\ A_{in}\ THEN\ y\ is\ B_i \quad (i = 1, 2, \ldots, m) \quad (1)$$

Here, m is the number of fuzzy production rules, and n is the number of premise variables x_j. y is consequence variable. Let $\mu_{A_{ij}}$ and μ_{B_i} be membership functions of the fuzzy set A_{ij} and B_i, respectively.

For simplicity, we write “IF” and “THEN” parts in the rules by the following notation: $\mathcal{A}_i = (\mu_{A_{i1}}, \mu_{A_{i2}}, \ldots, \mu_{A_{in}})$ $(i = 1, 2, \ldots, m)$, $\mathcal{A} = (\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_m)$ and $\mathcal{B} = (\mu_{B_1}, \mu_{B_2}, \ldots, \mu_{B_m})$. Then, the IF-THEN type fuzzy control rules above is called a fuzzy controller, and is denoted by $(\mathcal{A}, \mathcal{B})$. In the rules, the tuple of premise variable $x = x_1, x_2, \ldots, x_n$ is called an input information given to the fuzzy controller $(\mathcal{A}, \mathcal{B})$, and y is called an control variable.

2.2 Approximate Reasoning

It is confirmed by Ohta that a computational complexity of area method is smaller than gravity method, and there is hardly a difference of the control performance between the two from simulation of the auto cruise control [11]. In the following, the approximate reasoning method with area defuzzification method is introduced. If given the input information $x^* = ({x_1}^*, {x_2}^*, \ldots, {x_n}^*)$ to the fuzzy controller $(\mathcal{A}, \mathcal{B})$, the procedure for inference is summarized as follows:

Procedure 1. The strength of each rule is calculated by

$$\alpha_{\mathcal{A}_i}(x^*) = \prod_{j=1}^{n} \mu_{A_{ij}}({x_j}^*) \quad (i = 1, 2, \ldots, m).$$

Procedure 2. The control output of the each rule is calculated by

$$\beta_{\mathcal{A}_i B_i}(x^*, y) = \alpha_{\mathcal{A}_i}(x^*) \cdot \mu_{B_i}(y) \quad (i = 1, 2, \ldots, m),$$

where $\{\cdot\}$ indicates multiplication.

Procedure 3. The defuzzified value y_i^* of the fuzzy set in consequent part of i-th rule is given by

$$y_i^* = \frac{\int y\mu_{B_i}(y)dy}{\int \mu_{B_i}(y)dy} \quad (i = 1, 2, \ldots, m).$$

Procedure 4. Defuzzification stage.

$$\rho_{\mathcal{A}\mathcal{B}}(x^*) = \frac{\sum_{i=1}^m y_i^* S_{\mathcal{A}_i B_i}(x^*)}{\sum_{i=1}^m S_{\mathcal{A}_i B_i}(x^*)}, \quad \text{where} \quad S_{\mathcal{A}_i B_i}(x^*) = \int \beta_{\mathcal{A}_i B_i}(x^*, y)dy.$$

Since these calculations are depend on the membership functions, the subscripts $A_{ij}, \mathcal{A}_i, \mathcal{A}, B_i$, and $\mathcal{B}$ are put on α, β, S and ρ.

2.3 Fuzzy Feedback Control

Let $f(y, v) : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$ be a nonlinear vector valued function which is Lipschitz continuous. In addition, assume that there exists a constant $M_f > 0$ such that $\|f(y, v)\| \leq M_f (\|y\| + |v| + 1)$ for all $(y, v) \in \mathbb{R}^n \times \mathbb{R}$.

Consider a system given by the following state equation:

$$\dot{x}(t) = f(x(t), u(t)), \tag{2}$$

where $x(t)$ is the state and the control input $u(t)$ of the system is given by the state feedback $u(t) = \rho(x(t))$. Assume that the controllability is guaranteed in this system. For a sufficiently large $r > 0$, $B_r = \{x \in \mathbb{R}^n : \|x\| \leq r\}$ denotes a bounded set containing all possible initial states x_0 of the system. Let T be a sufficiently large final time. Then, we have

Proposition 1. [7] *Let $\rho : \mathbb{R}^n \to \mathbb{R}$ be a Lipschitz continuous function and $x_0 \in B_r$. Then, the state equation*

$$\dot{x}(t) = f(x(t), \rho(x(t))) \tag{3}$$

has a unique solution $x(t, x_0, \rho)$ on $[0, T]$ with the initial condition $x(0) = x_0$ such that the mapping $(t, x_0) \in [0, T] \times B_r \mapsto x(t, x_0, \rho)$ is continuous.

For any $r_2 > 0$, put

$$\Phi = \{\rho : \mathbb{R}^n \to \mathbb{R} : \textit{Lipschitz continuous}, \sup_{u \in \mathbb{R}^n} |\rho(u)| \leq r_2\}. \tag{4}$$

Then, the following (a) *and* (b) *hold.*

(a) *For any $t \in [0, T], x_0 \in B_r$ and $\rho \in \Phi$, $\|x(t, x_0, \rho)\| \leq r_1$, where*

$$r_1 = e^{M_f T} r + (e^{M_f T} - 1)(r_2 + 1). \tag{5}$$

(b) *Let $\rho_1, \rho_2 \in \Phi$. Then, for any $t \in [0, T]$ and $x_0 \in B_r$,*

$$\|x(t, x_0, \rho_1) - x(t, x_0, \rho_2)\| \leq \frac{e^{L_f (1 + L_{\rho_1})t} - 1}{1 + L_{\rho_1}} \sup_{u \in [-r_1, r_1]^n} |\rho_1(u) - \rho_2(u)|, \tag{6}$$

where L_f and L_{ρ_1} are the Lipschitz constants of f and ρ_1.

Let $r_2 \in \mathbb{R}$ be a positive constant, then r_1 is obtained by (5) in the proposition 1. In this study, for any fuzzy controller $(\mathcal{A}, \mathcal{B})$, the feedback function

$$\rho((x(t)) = \rho_{\mathcal{AB}}(x) : [-r_1, r_1]^n \to \mathbb{R}$$

on the basis of the rules (1) is defined by the previous approximate reasoning using area method, where $\rho_{\mathcal{AB}}(x) = \rho_{\mathcal{AB}}(x(t))$ is the amount of operation from the fuzzy controller $(\mathcal{A}, \mathcal{B})$ for the input information $x(t)$.

2.4 The Set of Membership Functions on L^∞ Space

Let $C[-r_1, r_1]$ be the Banach space of all continuous real functions on $[-r_1, r_1]$ with the norm $\|\mu\| = \max_{x\in[-r_1,r_1]} |\mu(x)|$. Denote by $L^1[-r_2, r_2]$ the Banach space of all Lebesgue measurable real functions μ on $[-r_2, r_2]$ such that

$$\int_{-r_2}^{r_2} |\mu(x)|dx < \infty.$$

We also denote by $L^\infty[-r_2, r_2]$ the Banach space of all Lebesgue measurable, essentially bounded real functions on $[-r_2, r_2]$. Let $\Delta_{ij} > 0$ $(i = 1, 2, \ldots, m;\ j = 1, 2, \ldots, n)$. We consider the following two sets of fuzzy membership functions.

$$F_{\Delta_{ij}} = \{\mu \in C[-r_1, r_1] : 0 \le \mu(x) \le 1 \ \text{ for } \forall x \in [-r_1, r_1],$$
$$|\mu(x) - \mu(x')| \le \Delta_{ij}|x - x'| \text{ for } \ \forall x, x' \in [-r_1, r_1]\}$$

and

$$G = \{\mu \in L^\infty[-r_2, r_2] : 0 \le \mu(x) \le 1 \text{ a.e. } x \in [-r_2, r_2]\}.$$

The set $F_{\Delta_{ij}}$, which is more restrictive than G, contains triangular, trapezoidal (π-type), bell-shaped, Z-type, and S-type fuzzy membership functions with gradients less than positive value Δ_{ij}. Consequently, if $\Delta_{ij} > 0$ is taken large enough, $F_{\Delta_{ij}}$ contains almost all fuzzy membership functions which are used in practical applications. In this study, we shall assume that the fuzzy membership functions $\mu_{A_{ij}}$ in “IF” parts of the rules (1) belong to the set $F_{\Delta_{ij}}$. On the other hand, we shall also assume that the fuzzy membership functions μ_{B_i} in “THEN” parts of the rules (1) belong to the set of discontinuous functions G [12].

Put

$$\mathcal{F} = \prod_{i=1}^{m} \left\{ \prod_{j=1}^{n} F_{\Delta_{ij}} \right\} \times G^m, \tag{7}$$

where G^m denotes the m times Cartesian product of G. Then, every element $(\mathcal{A}, \mathcal{B})$ of $\mathcal{F}$ is a fuzzy controller given by the fuzzy control rules (1).

2.5 Admissible Fuzzy Controller

The reasoning calculation and defuzzification are denoted as composite function through the inference procedure from 1 to 4 on the set of premise valuable

$[-r_1, r_1]^n$, and depend on the set of membership function. To avoid making the denominator of the expression in the procedure 3 and defuzzification stage equal to 0, for any $\delta > 0$, and $\varepsilon > 0$, we consider the subset

$$\mathcal{F}_{\delta\varepsilon} = \Bigg\{ (\mathcal{A}, \mathcal{B}) \in \mathcal{F} \ : \ \forall x \in [-r_1, r_1]^n, \sum_{i=1}^{m} S_{\mathcal{A}_i B_i}(x) \geq \varepsilon,$$

$$\forall i = 1, 2, \ldots, m, \int_{-r_2}^{r_2} \mu_{B_i}(y) dy \geq \delta \Bigg\}. \tag{8}$$

This is a slight modification of $\mathcal{F}$ by (7). If δ and ε are taken small enough, it is possible to consider $\mathcal{F} = \mathcal{F}_{\delta\varepsilon}$ for practical applications.

3 Continuity of Defuzzification as Functional

In this section, the continuity of approximate reasoning as functional on the set of membership functions $\mathcal{F}_{\delta\varepsilon}$ is shown. It is already shown that the calculations in the procedure 1 are continuous [8]. That is, if for each $i = 1, 2, \ldots, m$,

$$\mu_{A_{ij}^k} \to \mu_{A_{ij}} \ (i = 1, 2, \ldots, m; \ j = 1, 2, \ldots, n)$$

for $k \to \infty$ implies

$$\|\alpha_{\mathcal{A}_i^k} - \alpha_{\mathcal{A}_i}\|_\infty = \sup_{x \in [-r_1, r_1]^n} |\alpha_{\mathcal{A}_i^k}(x) - \alpha_{\mathcal{A}_i}(x)| \to 0.$$

Assume that a sequence $(\mathcal{A}^k, \mathcal{B}^k) \subset \mathcal{F}_{\delta\varepsilon}$ converges to $(\mathcal{A}, \mathcal{B})$ for the product topology if and only if, for each $i = 1, 2, \ldots, m$,

$$\|\alpha_{\mathcal{A}_i^k} - \alpha_{\mathcal{A}_i}\|_\infty \to 0$$

and

$$\mu_{B_i^k} \to \mu_{B_i} \text{ for the weak topology } \sigma(L^\infty, L^1). \tag{9}$$

Noting that for all $i = 1, 2, \ldots, m$, $\displaystyle\int_{-r_2}^{r_2} \mu_{B_i}(y) dy \geq \delta$ by (8) and the definition of membership function, we have

$$|y_i^{*k} - y_i^*| \leq \frac{r_2}{\delta^2} \Bigg(2 \left| \int_{-r_2}^{r_2} y \mu_{B_i^k}(y) dy - \int_{-r_2}^{r_2} y \mu_{B_i}(y) dy \right|$$

$$+ r_2 \left| \int_{-r_2}^{r_2} \mu_{B_i^k}(y) dy - \int_{-r_2}^{r_2} \mu_{B_i}(y) dy \right| \Bigg). \tag{10}$$

This means that the defuzzified value y_i^* is continuous on $\mathcal{F}_{\delta\varepsilon}$ by (9). Noting that $|\alpha_{\mathcal{A}_i^k}(x)| \leq 1$ $(\forall x \in [-r_1, r_1]^n, \forall i = 1, 2, \ldots, m)$, it is easy to show that

$$|S_{\mathcal{A}_i^k B_i^k}(x) - S_{\mathcal{A}_i B_i}(x)| \leq \left| \int_{-r_2}^{r_2} \mu_{B_i^k}(y) dy - \int_{-r_2}^{r_2} \mu_{B_i}(y) dy \right| + 2r_2 \|\alpha_{\mathcal{A}_i^k} - \alpha_{\mathcal{A}_i}\|_\infty. \tag{11}$$

It follows from routine calculation that

$$|\rho_{\mathcal{A}^k\mathcal{B}^k}(x)-\rho_{\mathcal{A}\mathcal{B}}(x)|\le\frac{4m{r_2}^2}{\varepsilon^2}\left(\sum_{i=1}^m|y_i^{*k}-y_i^*|+\sum_{i=1}^m|S_{\mathcal{A}_i^kB_i^k}(x)-S_{\mathcal{A}_iB_i}(x)|\right)$$

by noting that $\sum_{i=1}^m S_{\mathcal{A}_i^kB_i^k}(x)\ge\varepsilon$. Hence, $(\mathcal{A}^k,\mathcal{B}^k)\to(\mathcal{A},\mathcal{B})$ implies that

$$||\rho_{\mathcal{A}^k\mathcal{B}^k}-\rho_{\mathcal{A}\mathcal{B}}||_\infty=\sup_{x\in[-r_1,r_1]^n}|\rho_{\mathcal{A}^k\mathcal{B}^k}(x)-\rho_{\mathcal{A}\mathcal{B}}(x)|\to 0$$

for $k\to\infty$ by (10) and (11). It is easy to lead that the functional ρ is continuous on $\mathcal{F}_{\delta\varepsilon}$ from above inequalities. Therefore the continuity of the area method on $\mathcal{F}_{\delta\varepsilon}$ is obtained.

4 Lipschitz Continuity and Existence of Unique Solution of the State Equation

In this section, Lipschitz continuity of the defuzzification as the composite function on the premise variable $[-r_1,r_1]$ is shown. Lipschitz condition is applied to the existence of unique solution of the state equation (3). For all $i=1,2,\dots,m$, the following mapping $\alpha_{\mathcal{A}_i}$ are Lipschitz continuous on the space of premise variables $[-r_1,r_1]^n$ [7] [8]. That is, for $x,x'\in[-r_1,r_1]^n$

$$|\alpha_{\mathcal{A}_i}(x)-\alpha_{\mathcal{A}_i}(x')|\le\Delta_{\alpha_i}\|x-x'\|$$

and

$$|\beta_{\mathcal{A}_iB_i}(x,y)-\beta_{\mathcal{A}_iB_i}(x',y)|\le\Delta_{\alpha_i}\|x-x'\|,$$

where Δ_{α_i} is Lipschitz constant of $\alpha_{\mathcal{A}_i}$ $(i=1,2,\dots,m)$. Then we have

$$|S_{\mathcal{A}_iB_i}(x)-S_{\mathcal{A}_iB_i}(x')|\le 2r_2\Delta_{\alpha_i}\|x-x'\|.$$

Moreover, it follows from above Lipschitz continuity of $S_{\mathcal{A}_iB_i}$ that

$$|\rho_{\mathcal{A}\mathcal{B}}(x)-\rho_{\mathcal{A}\mathcal{B}}(x')|\le\frac{8{r_2}^3}{\delta^2}\sum_{i=1}^m\Delta_{\alpha_i}\|x-x'\|,\qquad(12)$$

by noting that $|y_i^*|\le r_2$. Therefore defuzzification by area method is Lipschitz continuous on the space of premise valuables $[-r_1,r_1]^n$.

Let $(\mathcal{A},\mathcal{B})\in\mathcal{F}_{\delta\varepsilon}$. Then it follows from Lipschitz continuity above that the $\rho_{\mathcal{A}\mathcal{B}}$ is Lipschitz continuous on $\mathbb{R}^n$ and satisfies $\sup_{u\in\mathbb{R}^n}|\rho_{\mathcal{A}\mathcal{B}}(u)|\le r_2$. Therefore, by the proposition 1 the state equation (3) for the feedback law $\rho_{\mathcal{A}\mathcal{B}}$ has a unique solution $x(t,x_0,\rho_{\mathcal{A}\mathcal{B}})$ with the initial condition $x(0)=x_0$ [13]. The detail about the extension of domain is omitted.

5 Application to Optimal Control Problem

In this section, using an idea and framework mentioned in the previous section, the existence of optimal control based on fuzzy rules in the admissible fuzzy controller will be established.

The performance index of this control system for the feedback law ρ in the previous section is evaluated with the following integral performance function:

$$J = \int_{B_r} \int_0^T w(x(t,\zeta,\rho), \rho(x(t,\zeta,\rho)))dtd\zeta.$$

where $w : \mathrm{I\!R}^n \times \mathrm{I\!R} \to \mathrm{I\!R}$ is a positive continuous function. The following theorem guarantee the existence of a rule set which minimizes the previous function.

Theorem 1. *The mapping*

$$\mathcal{F}_{\delta\varepsilon} \ni (\mathcal{A},\mathcal{B}) \mapsto \int_{B_r} \int_0^T w(x(t,\zeta,\rho_{\mathcal{A}\mathcal{B}}), \rho_{\mathcal{A}\mathcal{B}}(x(t,\zeta,\rho_{\mathcal{A}\mathcal{B}})))dtd\zeta \tag{13}$$

has a minimum (*maximum*) *value on the compact space* $\mathcal{F}_{\delta\varepsilon}$ *defined by* (8).

Proof. It is sufficient to prove that compactness of $\mathcal{F}_{\delta\varepsilon}$ and the continuity of performance function J on $\mathcal{F}_{\delta\varepsilon}$ are obtained.

For each $i = 1, 2, \ldots, m$, $\mathcal{F}_{\Delta_{ij}}$ is a subset of $C[-r_1, r_1]$ which is the subspace of continuous function on $[-r_1, r_1]$. Then it is compact respect for uniform norm $\|\cdot\|_\infty$ by the Ascoli Arzela's theorem [14]. On the other, G is closed for $\sigma(L^\infty, L^1)$. Then it is a compact set respect for the weak topology [7]. Therefore, by the Tychonoff's theorem, $\mathcal{F}$ is compact respect for the product topology. It is easy to show that $\mathcal{F}_{\delta\varepsilon}$ is a closed subset of $\mathcal{F}$ and hence it is compact. Assume that $(\mathcal{A}^k, \mathcal{B}^k) \to (\mathcal{A}, \mathcal{B})$ in $\mathcal{F}_{\delta\varepsilon}$ and fix $(t,\zeta) \in [0,T] \times B_r$. Then it follows from the section 3 that

$$\lim_{k\to\infty} \sup_{x\in[-r_1,r_1]^n} |\rho_{\mathcal{A}^k\mathcal{B}^k}(x) - \rho_{\mathcal{A}\mathcal{B}}(x)| = 0. \tag{14}$$

Hence, by (b) of the proposition 1, we have

$$\lim_{k\to\infty} \|x(t,\zeta,\rho_{\mathcal{A}^k\mathcal{B}^k}) - x(t,\zeta,\rho_{\mathcal{A}\mathcal{B}})\| = 0. \tag{15}$$

Further, it follows from (14), (15) and (a) of the proposition 1 that

$$\lim_{k\to\infty} \rho_{\mathcal{A}^k\mathcal{B}^k}(x(t,\zeta,\rho_{\mathcal{A}^k\mathcal{B}^k})) = \rho_{\mathcal{A}\mathcal{B}}(x(t,\zeta,\rho_{\mathcal{A}\mathcal{B}})). \tag{16}$$

Noting that $w : \mathrm{I\!R}^n \times \mathrm{I\!R} \to \mathrm{I\!R}$ is positive and continuous, it follows from (15), (16) and the Lebesgue's dominated convergence theorem [15] that the mapping (13) is continuous on the compact space $\mathcal{F}_{\delta\varepsilon}$. Thus it has a minimum (maximum) value on $\mathcal{F}_{\delta\varepsilon}$, and the proof is complete.

6 Conclusion

In this paper, we analyzed the continuity of the area defuzzification and proved that there exists an optimal feedback control law in a nonlinear fuzzy feedback control system, in which the feedback laws are determined by IF-THEN type fuzzy rules. To select the proper defuzzification method, it is necessary to analyze the property of fuzzy approximate reasoning and the simulation of fuzzy control.

It is recognized that in various applications it could be a useful tool in analyzing the convergence of fuzzy control rules modified recursively.

References

1. Tanaka, K., Sugeno, M.: Stability Analysis of Fuzzy Systems and Construction Procedure for Lyapunov Functions. Transactions of the Japan Society of Mechanical Engineers (C) 58(550), 1766–1772 (1992)
2. Hojo, T., Terano, T., Masui, S.: Fuzzy Feedback Control Rules Based on Optimality. Journal of Japan Society for Fuzzy Theory and Systems 5(5), 1200–1211 (1993)
3. Diamond, P.: Stability and periodicity in fuzzy differential equations. IEEE Trans. Fuzzy Syst. 8(5), 583–590 (2000)
4. Furuhashi, T.: Stability Analysis of Fuzzy Control Systems Based on Symbolic Expression. Journal of Japan Society for Fuzzy Theory and Systems 14(4), 357–365 (2002)
5. Ishibuchi, H., Nii, M.: Generating Fuzzy Classification Rules from Trained Neural Networks. Journal of Japan Society for Fuzzy Theory and Systems 9(4), 512–524 (1997)
6. Nomura, H., Wakami, N.: A Method to Determine Fuzzy Inference Rules by a Genetic Algorithm. The Transactions of the Institute of Electronics, Information and Communication Engineers (A) J77-A(9), 1241–1249 (1994)
7. Mitsuishi, T., Kawabe, J., Wasaki, K., Shidama, Y.: Optimization of Fuzzy Feedback Control Determined by Product-Sum-Gravity Method. Journal of Nonlinear and Convex Analysis 1(2), 201–211 (2000)
8. Mitsuishi, T., Endou, N., Shidama, Y.: Continuity of Nakamori Fuzzy Model and Its Application to Optimal Feedback Control. In: Proc. IEEE International Conference on Systems, Man and Cybernetics, pp. 577–581 (2005)
9. Mizumoto, M.: Improvement of fuzzy control (II). In: Proc. 4th Fuzzy System Symposium, pp. 91–96 (1988)
10. Terano, T.: Practical Fuzzy Control Technology. IEICE, Tokyo (1991)
11. Ohta, N., Harata, Y., Hayakawa, K.: A Fuzzy Inference LSI for an Automotive Control. R&D Review of Toyota CRDL 30(2), 45–55 (1995)
12. Gonda, E., Miyata, H., Ohkita, M.: Self-Tuning of Fuzzy Rules with Different Types of MSFs. Journal of Japan Society for Fuzzy Theory and Intelligent Informatics 16(6), 540–550 (2004)
13. Miller, R.K., Michel, A.N.: Ordinary Differential Equations. Academic Press, New York (1982)
14. Riesz, F., Sz.-Nagy, B.: Functional Analysis. Dover Publications, New York (1990)
15. Dunford, N., Schwartz, J.T.: Linear Operators Part I: General Theory. John Wiley & Sons, New York (1988)

Agent-Based In-Store Simulator for Analyzing Customer Behaviors in a Super-Market

Takao Terano[1], Ariyuki Kishimoto[1], Toru Takahashi[1], Takashi Yamada[1],
and Masakazu Takahashi[2]

[1] Tokyo Institute of Technology, 4259-J2-52, Nagatsuda-Cho,
Midori-ku, Yokohama 226-8502, Japan
terano@dis.titech.ac.jp, kiseetoworld3188@gmail.com,
{toru,tyamada}@trn.dis.titech.ac.jp
[2] University of Tsukuba, 3-29-1, Otsuka, Bunkyo-ku, Tokyo 112-0012, Japan
masakazu@gssm.otsuka.tsukuba.ac.jp

Abstract. This paper presents an agent-based simulator to investigating customer walking flows and purchasing behaviors in a super market. So far, such investigations have cost very much to examine in real situations. The simulator enables us to carry out “virtual experiments” through changing various parameters of retail businesses and store operations. For this purpose, first we observe an actual retail store and analyze sales data. Then we develop the simulation model: Agent-Based In-Store Simulator (ABISS). Intensive experiments have revealed that the flow of customers, which is related to the sales, depends on the design of a store and that the places of in-store advertisement and recommendation system vary their sales.

Keywords: Multi-agent Based Simulation, Agent-Based Modeling, Customer Behaviors, Data Mining, Decision Support.

1 Introduction

Service management of retail business in a super-market requires investigations for store operations including the shop layout, sales promotion, and control of customer flows. So far, such investigations have cost very much to examine in real situations. In this study, we will carry out “virtual experiments” for the purpose through changing various parameters on retail businesses and store operations.

For the purpose, we have developed Agent-Based In-Store Simulator (ABISS), which enable us to examine the effects of changes of operations through detailed simulation settings. ABISS consists of 1) store layout information, which includes kinds of sales items, item category groups, and promotion panels, 2) customer agents, which have decision making functionality for personal behaviors and purchasing items, 3) ‘virtual’ recommendation functions for the customers, and 4) data bases on sales items, their stocks, and point-of-sales data for each customer.

The data of items, customers and layout information are implemented based on the ones of a real small sized store in Japan. However, the information is easily modified

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 244–251, 2009.

according to the changes of situations. Customer agents are able to walk around the virtual store and purchase items in the store according to their decisions and recommendations of the store. The virtual point-of-sales data are gathered and analyzed by data mining techniques during and/or after simulation studies. Therefore, using ABISS, we virtually but efficiently investigate the effects of changes of retail management and operation strategies.

This paper presents the basic principles and architectures of ABISS and experimental results of the simulation, which have revealed that 1) customer behaviors both in the current layout and in different modified ones with catalysts for circulation changes: 2) the effects of design changes of new floor layouts and placement of advertisements, and 3) effects of personal and/or group recommendations for sales promotions.

The rest of the paper is organized as follows: Section 2 discusses the background of the research and related work; Section 3 briefly explains the field study on the target super-market; Section 4 describes the basic principles and architecture of ABISS; Section 5 presents experimental results; and Section 6 gives some concluding remarks and future work.

2 Background and Related Work

There have been so many researches about consumer behaviors and their decision analysis in marketing science literature [1],[2]. They are based on statistical analysis of purchased items and profiles of consumers. Recently, several data mining techniques have been used for the analysis [3],[4]. They have been also utilized customers' movement data in a real shop using recent technologies of RFID tags and video camera tracking [5],[6], [7]. The data obtained in these investigation methods are intrinsically very noisy and it costs much to carry out the investigations.

We are conducting a research project to develop a decision support system to increase the service productivity on retail store management including ordinary super markets [8], [9]. In the current situation, however, to measure the individual customers' behaviors in real time is very difficult because of the measurement costs. Instead of carrying out real experiments, we have decide to develop a 'virtual' environment for the purpose. We employ agent-based modeling (ABM) to simulate consumer behaviors in a super-market.

Agent-based modeling is one of the cutting-edge methodologies to uncover social and economic behaviors of individuals [10], [11]. The "agent" is a software component, which models humans, groups, organizations, and societies. An agent contains internal states, decision and/or behavioral rules, and communication capability among them. Using ABM, we would like to uncover emergent behaviors as results of interactions of micro- and macro-links of individuals and their group behaviors. In this study, using bottom-up manner, we design customers as agents and a super-market as a physical and information environment. We set various micro level parameters of agents and their environment so that macro level emergent phenomena are coinciding with the real observable conditions. Also, because the ABM is implemented as a computer simulator, the explicability of the model is easily validated. Using the model with reasonable time and costs, we are able to evaluate conditions like plural floor layouts, sales promotion tactics, and marketing strategies.

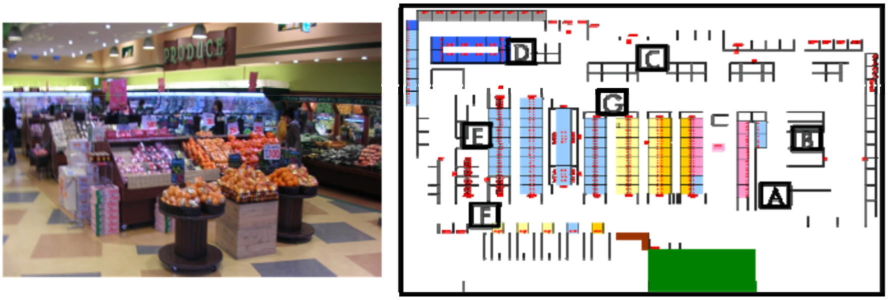
3 Field Study of the Target Store

Our objectives of ABISS are 1) to explore in-store customer behavior analysis to increase customer convection time and purchase points, and 2) to examine store layout changes and POP advertisement methods.

Before developing ABISS, we have carried out a detailed fieldwork on the targeting super-market, whose photo and the basic layout is shown in Figure 1. In the figure (b), marks A, B, …, F in (b) mean important areas for the analysis.

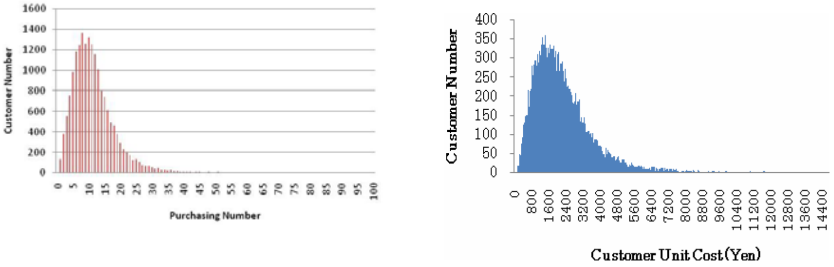
Interview survey with store management, we have obtained the following comments:

- There are many customers that move to the inner part of the store after having perused the main aisle which centers on the 3 fresh items.
- The time spent in the store is about 20 minutes for each customer.
- Modifying store layout and new item sales promotions requires a lot of time and money, and are difficult to execute.



(a) Photo of the Store (b) Current Floor Layout

Fig. 1. A Photo of the Target Store and the Current Floor Layout



(a) Customers vs. Number of Items (b) Customers vs. Purchased Costs

Fig. 2. Customer Distribution of Purchased Item Numbers and Purchased Costs

Through the field work, we have understood customer movement lines, and the factors to induce customer purchases in sales areas.

In addition, we analyzed the ID attached POS data (between 2007/8/31~2007/9/29, with total transactions totaling 51,109), we have analyzed the average customer interval, the average PI value every of sale area, the average unit price per item of every sale area, the customer distribution of every item's purchase score (Figure 2 (a)), and the customer distribution of every customer unit cost (Figure 2 (b)).

4 How ABISS Works

This section briefly describes the architecture of ABISS: design of the store space and its environment; agent decisions and movement.

4.1 Store Space Layout

The in-store space is displayed in a two-dimensional space with discrete cells. The customer agent is defined as the human body's circular radius of 0.3 meters, and one cell of the simulation space is defined as the human body circular diameter of 0.6m X 0.6m. In addition, the following objects exist in a cell.

Customer Agent represents an in-store customer .

Walls represent obstacle objects in movement of customers, such as an item shelf and a store wall.

Sales Promotion Objects are used to represent promotion objects such as advertisements inside the store are displayed.

Item Sale Area: $\boldsymbol{U = \{u_i \mid i = 0, , m\}}$ means a set of item positions of the item shelf. U also holds the respective average item unit cost $C = \{ci \mid i = 0, m\}$.

Entrance is an area where customers enter the store is displayed

Register is an area to carry out transaction processing .

4.2 Customer Agents

Customer agents move into the store while taking planned purchase items, which have been previously decided to be purchased. While moving, sale items within the vision range are happened to be unplanned purchases. Also, when sales promotion items such as advertisements are found, they are also nominated as unplanned purchases. When all purchases are made, the agent will move towards the register, and make the transactions. When the transactions are made, the in-store actions will be complete.

Each customer agent has the following parameters.

Number of planned purchase items: $\boldsymbol{N_{plan}}$ means the number of the planned purchase items before entering the store. N_{plan} is 30% of the buying score distribution.

Planned purchase item selection probability: $\boldsymbol{P_{plan}}$ means the probability of selecting a planned purchase item.

Planned purchase item list: $\boldsymbol{L_{plan}}$ represents the list of item sale areas which are planned purchases held before entering the store.

Estimate is the upper limit amount of items purchased by a customer .

Round schedule sale area selection probability: P_{round} is the probability to move round scheduled sale area.

Walking speed determines the speed a customer agent moves.

Visual purchasing probability : P_{watch} means the probability of unplanned purchase happening, when visiting or passing an item sales area.

Range of Vision represents how far agents can see sales promotion item objects.

Recognition Ratio is the possibility of recognizing the information of the sales items when sales promotions come into the range of vision.

Sales promotion purchasing probability : $P_{promotion}$ displays the probability to accept the advertisement of a sales promotion item.

Sales promotion purchase item sales area list : $L_{promotion}$ holds the sales area of the item by sales promotion and advertising.

Scheduled visit sale area list : L_{all} holds scheduled visit sale areas.

Shopping cart holds the items acquired from sales areas.

Total payment amount displays the total amount of the items which are going to be purchased.

4.3 Behaviors of Customer Agents

When the agent enters the store, planned purchase items are not influenced by the store's quality or interior conditions, and planned purchase items and the store interior determines the agent's movement path. We assume that all items which enter the cart will be purchased. The pseudo code of the customer agent conduct rule is shown in Figure 3.

```
//purchasing_behavior
 Agent enters the store as UnplannedCustomer
 Determine purchased items with planned purchased probability Pplan
 Lplan :-  SalesArea on purchased items
 If total payment > possessed money then CustomerAgent becomes PlannedCustomer
 Determine Lvisit, where PlannedCustomer finds items with probability Pround
 Lall :- Lplan U Lvisit
 Loop
  call in-store_movement
  Lall :- Lall - SalesArea
  If(Lall is empty) {call payment; exit}
  If (total payment = possessed money) {call payment; exit}
 End Loop

//in-store_movement
Loop
 While (Agent is not at SalesArea in Lall) call walk_around
 /*planned purchase
 If(Agent arrives at SalesArea in Lall) then call stay_to_purchase
 /*unplanned purchase
 If(Agent passes SaleAea) then  call visual_purchase
 If(Agent meets SalesPromotionMedia) then call promotion_purchase
End Loop
```

Fig. 3. Pseudo-Code for the Customer Behaviors

5 Experimental Results

Based on the model Section 4, we have conducted intensive simulation experiments. First, to validate the simulator, we have run the simulation at the original layout and data. We have tuned the parameters up and the results have coincided with the actual data. Second, we have changed the floor layout. Third, we have added advertisements tactics to them. We describe the second and third results in this section.

We have prepared 16 kinds of fresh product department layouts which are changed from the original layout, and have conducted three 43,2000-step (representing the store's opening time) simulations. The experiment settings are summarized in Table 1. We have analyzed how the changes in floor layout affected the behavior of customers.

Table 1. Experiment Settings on Floor Layout Design

Layout	Basic Layout
	Vegetable Layouts 1~9
	Meat Layout
	Fish Layout
	Vegetable-Meat/Fish Layout
	3 Fresh Products Layout
	Divided Vegetable Layout
Simulation Time	9：00 ~ 21：00 (43,200 simulation steps)
No. of Simulation Runs	3 times

- **Basic layout:** Same layout as the surveyed super-market
- **Vegetable layout 1-9:** The fruit/vegetable department layouts are changed
- **Meat layout:** The meat, processed meat, fish sausage, and meat-related spice & condiment department was switched with the prepared foods & bread department in the back-left corner of the store.
- **Fish layout:** The fish & dried/salted foods department was switched with the prepared foods & bread department in the back-left corner of the store.
- **Vegetable-Meat/Fish layout:** The fruit/vegetable department and the vegetable-related dressing /juice departments were switched with the meat & fish-related department.
- **3 Fresh Products Layout: V**egetables, meat, and fish-related items departments have been moved to the back-left corner of the store.
- **Divided vegetable layout:** The vegetable/fruit-related department has been split into two departments.

The graphs for circulation and number of visual purchases per customer in all layouts are displayed in Figures 4 and 5.

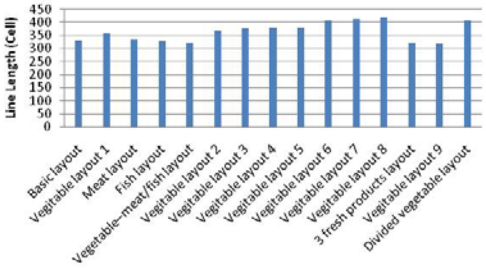


Fig. 4. Circulation Per Customer (All Operating Hours)

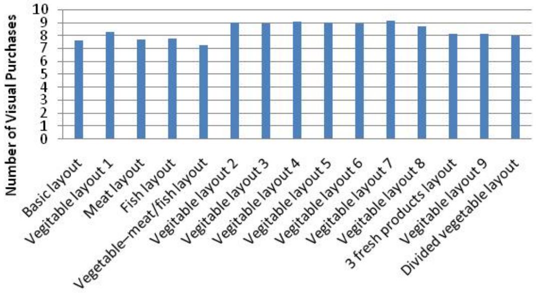
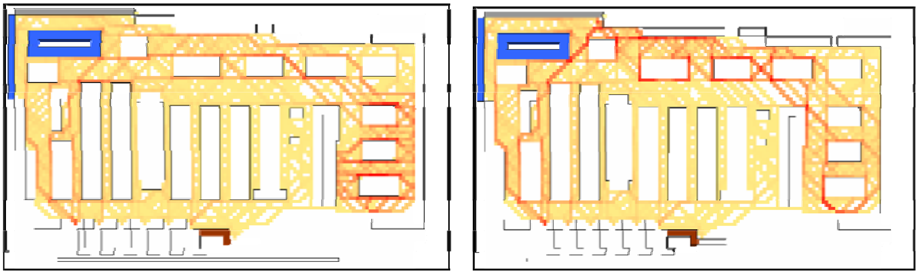


Fig. 5. Visual Purchases Per Customer (All Operating Hours)

In each experiment, we have observed various simulation snapshots to interpret the results, however, because of the space limit, we have skipped detailed ones in the paper.

About the third experiments, Figure 6 depicts samples of simulated footprints of the original layout and the effects of personal and/or group recommendations for sales promotions. In the experiments, promotion posters are assumed to be used. The darker meshes mean the more customer move the place. We have found that the places of the promotion information will remarkably affect the customers' behaviors in a store.



(a) Basic Layout (b) Case with Promotion Recommendation

Fig. 6. Results of Customers' Footprints in the Two Different Floor Layouts

6 Concluding Remarks

This paper has presented Agent-Based In-Store Simulator (ABISS) for analyzing customer behaviors in a super-market, which aims at developing a marketing decision support system for retail management. We have described the background, basic architecture, and experimental results of ABISS. Before executing simulation studies, we have carried out intensive field survey and POS data analysis to uncover the current situations of the targeted super-market. The experimental results have suggested the applicability and effectiveness of ABM to carry out "virtual experiments" in the marketing decision support domains with various parameters.

Our future work includes 1) investigation of real and simulated data on customer behaviors, 2) development of decision support functions for store managers, 3) analysis regarding customer information and product categories, and 4) grounding the simulation results with real customer behavior data. These work will require practical experiments and further survey studies.

References

1. Guadagni, P.M., Little, J.D.C.: A Logit Model of Brand Choice, Calibrated on Scanner Data. Marketing Science 2, 203–238 (1983)
2. Gupta, S.: Impact of Sales Promotion on When, What, and How Much to Buy. Journal of Marketing Research 25, 342–355 (1988)
3. Yada, K., Washio, T., Motoda, H.: Consumer Behavior Analysis by Graph Mining Technique. New Mathematics and Natural Computation 2, 59–68 (2006)
4. Yada, K., Ip, E., Katoh, N.: Is This Brand Ephemeral? A Multivariate Tree-Based Decision Analysis of New Product Sustainability, Decision Support Systems 44, 223–234 (2007)
5. Sorensen, H.: The Science of Shopping. Marketing Research 15, 30–35 (2003)
6. Larson, J.S., Bradlow, E.T., Fader, P.S.: An Exploratory Look at Supermarket Shopping Paths. International Journal of Research in Marketing 22, 395–414 (2005)
7. Sato, H., Kubo, M., Namatame, A.: A Method to Translate Customers' Actions in Store into the Answers of Questionnaire for Conjoint Analysis. In: Terano, T., Kita, H., Takahashi, S., Deguchi, H. (eds.) Agent-Based Approaches in Economic and Social Complex Systems V. Springer ABSS, vol. 6, pp. 157–168 (2008)
8. Terano, T., Yamada, T., Takahashi, M., Nakao, T., Kishimoto, A., Yokokawa, M.: Toward Agent-Based Simulation on Marketing Behaviors for Service Science Studies. In: 3rd International Nonlinear Sciences Conference (INSC 2008) -Application in Behavioral, Social & Life Sciences, Chuo University, Faculty of Commerce, pp. 15–16 (2008)
9. Takahashi, M., Nakao, T., Tsuda, K., Terano, T.: Generating Dual-Directed Recommendation Information from Point-of-Sales Data of a Supermarket. In: Lovrek, I., Howlett, R.J., Jain, L.C. (eds.) KES 2008, Part II. LNCS (LNAI), vol. 5178, pp. 1010–1017. Springer, Heidelberg (2008)
10. Terano, T.: Exploring the Vast Parameter Space of Multi-Agent Based Simulation. In: Antunes, L., Takadama, K. (eds.) MABS 2006. LNCS (LNAI), vol. 4442, pp. 1–14. Springer, Heidelberg (2007)
11. Terano, T.: Beyond the KISS Principle for Agent-Based Social Simulation. Journal of Socio-Informatics 1(2), 175–187 (2008)

Detecting Temporal Patterns of Importance Indices about Technical Phrases

Hidenao Abe and Shusaku Tsumoto

Department of Medical Informatics, Shimane University, School of Medicine
89-1 Enya-cho, Izumo, Shimane 693-8501, Japan
abe@med.shimane-u.ac.jp, tsumoto@computer.org

Abstract. In text mining, importance indices of terms such as simple frequency, document frequency including the terms, and tf-idf of the terms, play a key role for finding valuable patterns in documents. As for the documents, they are often published daily, monthly, annually, and irregularly for each purpose. Although the purposes of each set of documents are not changed, roles of terms and the relationship among them in the documents change temporally. In order to detect such temporal changes, we decomposed the process into three sub-processes: automatic term extraction, importance index calculation, and temporal trend detection. On the basis of the consideration, we propose a method for detecting temporal trends of technical terms based on importance indices and clustering methods. By focusing on technical phrases, we carried out an experimentation to detect emergent and subsiding trends in a set of research document. The result shows that our method determined the temporal trends of technical phrases related to finding of patterns for innovations of research topics.

Keywords: Text Mining, Trend Detection, TF-IDF, Jaccard Coefficient, Linear Regression.

1 Introduction

In recent years, developing information systems in every field such as enterprise, academic and medical organizations, and stored data have increased year by year. Accumulation is advanced to document data by not the exception but various fields. Especially, the document data gives valuable findings not only for domain experts in headquarter sections but also for novice users about particular domains such as day traders, news readers and so forth. In this situation, detecting a new phrases and terms has been very important. In order to realize the detection, emergent term detection (ETD) methods have been developed [1,2].

However, because the frequency of the word was used in earlier methods, detection was difficult as long as the word that became an object did not appear. In addition, almost of the conventional methods are not consider nature of terms and importance indices separately. This causes difficulties of text mining applications such as limitations for extensionality of time direction, time consuming

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 252–258, 2009.

post-processing, and other generality expansions. By considering the problem, we focus on temporal changes of importance indices of phrases and changes of them. The temporal changes of the importance indices of extracted phrases paid to attention so that the specialist may recognize an emergent terms or/and such fields in this research.

In this paper, we propose a method to detect trends of phrases by combining term extraction methods, importance indices of the terms and trend analysis methods in Section 2. Then, taking as an example of the titles and abstracts of IEEE International Conference of Data Mining (ICDM)[1], we show the comparison of changes in two importance indices in Section 3. Finally, in Section 4, we summarize this paper, and describe our future work.

2 A Method to Detect Trends of Importance Indices of Automatically Extracted Terms

In this section, we describe a method to detect some temporal trends of technical terms by using multiple important indices consisting of the following three subprocesses:

1. Technical term extraction in a corpus
2. Importance indices calculation
3. Trend detection

There are some conventional studies to extract technical terms in a corpus based on each particular importance index[2]. Although these methods calculate each index to extract technical terms, the information of the importance of each term are lost by cutting off the information with a threshold value. We suggest separating term determination and trends detection based on importance indices. By separating these phases, we can calculate multiple kinds of importance indices. Subsequently, to the dataset, we can apply many kinds of temporal analysis methods based on statistical analysis, clustering and machine learning algorithms. The overview of this method is illustrated in Figure 1.

Firstly, the system determines terms in a given corpus. There are two reasons why we introduced term extraction methods before calculating importance indices. One is that the costs to build up a dictionary for each particular domain are very expensive task. The other is the need to detect new concepts in a given temporal corpus. Especially, a new concept is often described in the document for which the character is needed at the right time in using the combination of existing words. By considering above reasons, we applied a term extraction method based on adjacent frequency of compound nouns. The method extracts the technical terms by using the following values for each candidates CN:

$$FLR(CN) = f(CN) \times (\prod_{i=1}^{L}(FL(N_i)+1)(FR(N_i)+1))^{\frac{1}{L}} > 1.0$$

[1] http://www.cs.uvm.edu/~icdm/

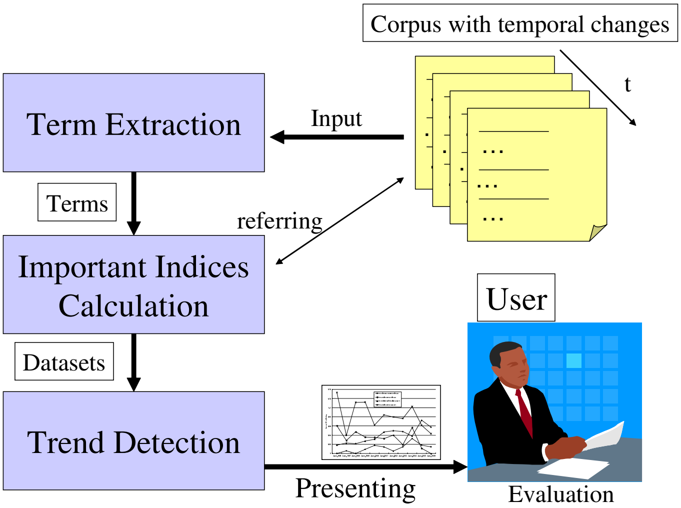


Fig. 1. An overview of the proposed remarkable temporal trend detection method

Where $f(CN)$ means frequency of the term CN. Similarly, $FL(N_i)$, and $FR(N_i)$ mean frequencies of the right and left of the object noun N_i. In order to determine terms in this part, we can also use the other term extraction methods and terms/keywords from users.

After determining terms in the given corpus, the system calculates multiple importance indices of the terms for the documents in each period. As for importance indices of words and phrases in a corpus, there are some well known indices. Term frequency divided by inversed document frequency (tf-idf) is one of the popular indices to measure the importance of the terms. The definition of tf-idf is shown in the following:

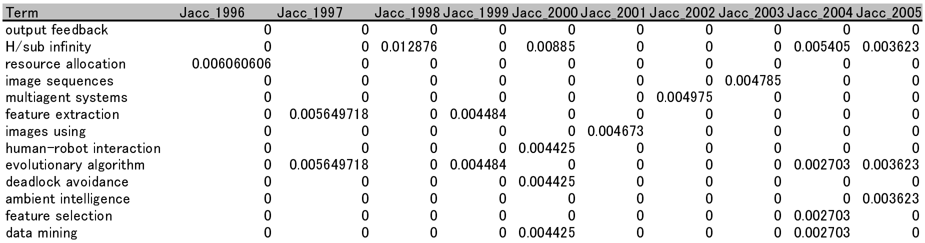
$$TFIDF(t) = \frac{TF(t)}{log_e \frac{|D|}{DF(t)}}$$

Where $TF(t)$ means the frequency of each term t in the corpus with $|D|$ documents. And $DF(t)$ means the frequency of documents containing w_i, which are the words included in the term t. As another importance index, we used Jaccard's matching coefficient[3][2] calculated as the following:

$$Jaccard(t) = \frac{h(w_1, w_2, ..., w_L)}{h(w_1)h(w_2)...h(w_L)}$$

Where $h(w_i)$ means the number of hit documents in the corpus to the word w_i. Each Jaccard coefficient value shows strength of co-occurrence of multiple words as an importance of the terms in the given corpus. We can also assume

[2] Here after, we call just "Jaccard coefficient".



Term	Jacc_1996	Jacc_1997	Jacc_1998	Jacc_1999	Jacc_2000	Jacc_2001	Jacc_2002	Jacc_2003	Jacc_2004	Jacc_2005
output feedback	0	0	0	0	0	0	0	0	0	0
H/sub infinity	0	0	0.012876	0	0.00885	0	0	0	0.005405	0.003623
resource allocation	0.006060606	0	0	0	0	0	0	0	0	0
image sequences	0	0	0	0	0	0	0	0.004785	0	0
multiagent systems	0	0	0	0	0	0	0.004975	0	0	0
feature extraction	0	0.005649718	0	0.004484	0	0	0	0	0	0
images using	0	0	0	0	0	0.004673	0	0	0	0
human-robot interaction	0	0	0	0	0.004425	0	0	0	0	0
evolutionary algorithm	0	0.005649718	0	0.004484	0	0	0	0	0.002703	0.003623
deadlock avoidance	0	0	0	0	0.004425	0	0	0	0	0
ambient intelligence	0	0	0	0	0	0	0	0	0	0.003623
feature selection	0	0	0	0	0	0	0	0	0.002703	0
data mining	0	0	0	0	0.004425	0	0	0	0.002703	0

Fig. 2. Example of the dataset consisting of an importance index

the degrees of co-occurrence such as the χ^2 statistics to the terms consisting of multiple words as the importance indices in our method.

In our method, we propose treating these indices explicitly as a temporal dataset. Figure 2 shows an example of the dataset consisting of an importance index for the years.

Then, the method provides the choice of some adequate trend extraction method such as linear regression analysis, clustering and so forth to the datasets. In the following case study, we applied the linear regression analysis technique in order to detect the degree of existing trends based on the two importance indices. The degree of each term t calculated as the following:

$$Deg(t) = \frac{\sum_{i=1}^{M}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{M}(x_i - \bar{x})},$$

where $\bar{x}$ means the average of the period, and $\bar{y}$ means the average of each importance index for the period $1 \leq i \leq M$. At the same time, we also calculated the intercept $Int(t)$ of each term t as the following:

$$Int(t) = \bar{y} - Deg(t)\bar{x}.$$

3 Experiment: Detecting Remarkable Trends of Technical Phrases in a Temporal Corpus

In this experiment, we show the results of detecting two trends by using the method described in Section 2. As the input of temporal documents, annual sets of titles and abstracts of ICDM conference from 2002 to 2008 are taken.

In these corpus, we determined technical terms by using the term extraction method[4][3] for each entire corpus.

Subsequently, tf-idf and Jaccard coefficient values are calculated for each term to the annual documents on each corpus. To the datasets consisting of temporal values of the important indices, we applied linear regression to detect the following two trends of the phrases: Emergent and Subsiding.

[3] The implementation is called Gensen, distributed in `http://gensen.dl.itc.u-tokyo.ac.jp/termextract.html` (in Japanese).

3.1 Extracting Technical Terms

As for the documents, we assumed each title and abstract of the article as one document. Then, we did not use any stemming technique, because we want to consider detailed difference of the terms.

Table 1 shows the description of the abstracts and titles of ICDM conferences from 2002 to 2008.

Table 1. Description of the ICDM corpus

Year	Abstract		Title	
	#documents	#words	#documents	#words
2002	112	18,916	112	960
2003	125	19,068	125	1,040
2004	106	15,985	106	840
2005	141	20,831	141	1,153
2006	152	24,217	152	1,307
2007	101	16,143	101	782
2008	144	22,971	144	1,136
TOTAL	881	138,131	881	7,218

We applied the term extraction method to all of the abstracts and the titles of the seven years' ICDM conferences. From of the entire abstracts for the seven years, the method extracted 21,599 terms.

As same as to the abstracts, 1,912 terms are extracted in the entire titles for the seven years.

3.2 Results of the Automatically Extracted Terms

By using the degree and the intercept of each term, we tried to determine the following two trends:

- Emergent
  - sorting the degree with ascending order
  - sorting the intercept with descending order
- Subsiding
  - sorting the degree with descending order
  - sorting the intercept with ascending order

Table2 shows the top ten phrases extracted in the abstracts of ICDM having the two trends based on the degrees $Deg(t)$ and the intercepts $Int(t)$ of every phrase t. The degrees and the intercepts are calculated for the two importance indices. As same as the phrases from the abstracts, Table3 shows the top ten phrases extracted in the titles of ICDM.

On each set of documents, our method identified similar phrases both of the two importance indices for each temporal trend. "Social Network" shows the most typical emergent trend with both of the importance indices. This indicates that both of the standardized frequency of the phrases and the characteristic

Table 2. Top 10 phrases with two trends based on tf-idf and Jaccard coefficient values in the abstracts of ICDM

Rank	Emergent						Subsiding					
	tf-idf			Jaccard Coefficient			tf-idf			Jaccard Coefficient		
	t	Deg(t)	Int(t)	t	Deg(t)	Int(t)	t	Deg(t)	Int(t)	t	Deg(t)	Int(t)
1	Experimental results	0.0050	0.0182	collaborative filtering	0.077	0.284	data mining	−0.021	0.222	association rules	−0.055	0.457
2	social network	0.0032	−0.0020	upper bound	0.065	0.089	association rules	−0.011	0.057	association rule	−0.053	0.421
3	text mining	0.0028	−0.0014	social networks	0.054	−0.035	experimental results	−0.009	0.072	web pages	−0.043	0.480
4	real world	0.0027	0.0040	social network	0.053	0.027	association rule	−0.006	0.033	nearest neighbor	−0.042	0.501
5	feature extraction	0.0024	−0.0029	gene expression	0.051	0.479	data sets	−0.006	0.084	frequent itemsets	−0.039	0.367
6	real-world data	0.0016	0.0054	matrix factorization	0.047	0.043	frequent itemsets	−0.005	0.039	naive Bayes	−0.034	0.266
7	background knowledge	0.0016	−0.0004	Support Vector Machines	0.045	0.143	clustering algorithm	−0.003	0.025	experimental results	−0.032	0.322
8	social networks	0.0014	−0.0015	anomaly detection	0.038	0.027	mining algorithm	−0.003	0.016	dynamic programming	−0.027	0.218
9	synthetic data	0.0014	0.0048	random walk	0.038	0.057	clustering algorithms	−0.002	0.025	outlier detection	−0.022	0.217
10	real datasets	0.0013	0.0016	computational cost	0.037	−0.024	association rule mining	−0.002	0.012	feature selection	−0.017	0.268

Table 3. Top 10 phrases with two trends based on tf-idf and Jaccard coefficient values in the titles of ICDM

Rank	Emergent						Subsiding					
	tf-idf			Jaccard Coefficient			tf-idf			Jaccard Coefficient		
	t	Deg(t)	Int(t)	t	Deg(t)	Int(t)	t	Deg(t)	Int(t)	t	Deg(t)	Int(t)
1	Data Streams	0.0010	0.0029	Collaborative Filtering	0.125	0.101	Association Rules	−0.00333	0.01638	Event Sequences	−0.107	0.607
2	Active Learning	0.0006	0.0000	Nonnegative Matrix Factorization	0.095	−0.048	Data Mining	−0.00111	0.01173	Association Rules	−0.089	0.583
3	Nonnegative Matrix Factorization	0.0005	−0.0007	Random Walk	0.095	0.048	Data Sets	−0.00104	0.00559	Decision Trees	−0.077	0.637
4	Collaborative Filtering	0.0005	0.0002	Dimension Reduction	0.089	0.018	Decision Trees	−0.00057	0.00404	Latent Semantic Indexing	−0.060	0.393
5	Text Categorization	0.0005	−0.0006	Hidden Markov	0.071	−0.071	Unsupervised Algorithm	−0.00048	0.00206	Experimental Evaluation	−0.054	0.375
6	Social Networks	0.0005	−0.0004	Belief Propagation	0.071	0.000	Web Page Classification	−0.00048	0.00206	Bayesian Network	−0.054	0.280
7	Dimension Reduction	0.0004	−0.0001	Taxonomic Research	0.071	0.071	Dimensional Data	−0.00040	0.00232	Document Categorization	−0.054	0.280
8	frequent itemsets	0.0004	−0.0002	Similarity Measure	0.065	−0.077	Time Series	−0.00038	0.00797	Fast Algorithm	−0.048	0.238
9	Document Clustering	0.0004	0.0003	Pairwise Constraints	0.051	−0.068	Latent Semantic Indexing	−0.00037	0.00227	Utility Itemsets	−0.045	0.313
10	Sequential Pattern Mining	0.0003	−0.0002	Link Prediction	0.050	0.021	data mining	−0.00037	0.00207	data mining	−0.042	0.244

combinations of the words included in the phrases has been focused as an important topic in these years by both of the reviewers and the authors.

Since Jaccard coefficient does not depend on the frequencies of the phrases, some different phrases such as “Matrix Factorization” are identified by this importance index. Looking at the shapes of the importance indices in these charts, tf-idf can detect more smooth temporal trends than those of Jaccard coefficient. The meaning of the difference between the importance indices should be clarified by connecting the other information in the future.

3.3 Discussion about the Empirical Results

As for the comparison between the phrases extracted by the term extraction method and the keyword specified by the authors, the emergent phrases and the subsiding phrases are almost same. Since research documents are usually written carefully, the result that there is no significant difference between the two sets of phrases is not surprising. However, especially for ill-written documents, the automatic term extraction will be helpful to identify un-expected phrases in an analyst mind. By cooperating with automatic term extraction methods, we can consider the temporal trends of both of expected and un-expected phrases in a temporal corpus.

The difference between the importance indices of phrases also should be discussed. As shown in Table 2 and Table 3, the behaviors of the two importance indices are different in spite of sorting with the same criteria calculated with the linear regression technique. We will improve the sorting result with introducing the other criteria to measure the fitness of each series of values to linear models such as coefficient of determination, statistical test values of the coefficient and other criteria of the linear regression model.

In this experiment, we only obtained the temporal trends from the view of un-supervised manner. Further experiments from the views of supervised and co-clustering will be needed to identify each functional characteristics of the importance indices.

4 Conclusion

In this paper, we proposed the method to detect trends of technical terms by focusing on the temporal changes of the importance indices. We implemented the method by combining the technical term extraction method, the two important indices, and linear regression analysis.

The case study shows that the temporal changes of the importance indices can detect the trend of each term, according to the degree of the values for each annual document. The emergent terms, which detected by a domain expert, are ranked as the terms with increasing degrees of the importance indices. Regarding to the result, our method can support to find out trends of terms in documents based on the temporal changes of the importance indices.

In the future, we will apply other term extraction methods, importance indices, and trend detection method. As for importance indices, we are planning to apply evaluation metrics of information retrieval studies, probability of occurrence of the terms, and statistics values of the terms. To extract the trends, we will introduce temporal pattern recognition methods, such as temporal clustering. Then, we will apply this framework to other documents from various domains.

References

1. Lent, B., Agrawal, R., Srikant, R.: Discovering trends in text databases, pp. 227–230. AAAI Press, Menlo Park (1997)
2. Kontostathis, A., Galitsky, L., Pottenger, W.M., Roy, S., Phelps, D.J.: A survey of emerging trend detection in textual data mining. A Comprehensive Survey of Text Mining (2003)
3. Anderberg, M.R.: Cluster Analysis for Applications. Monographs and Textbooks on Probability and Mathematical Statistics. Academic Press, Inc., New York (1973)
4. Nakagawa, H.: Automatic term recognition based on statistics of compound nouns. Terminology 6(2), 195–210 (2000)

Recommender System for Music CDs Using a Graph Partitioning Method

Takanobu Nakahara[1] and Hiroyuki Morita[2]

[1] Kansai University, 3-3-35 Yamate-cho, Suita-shi, Osaka, Japan
nakapara@ipcku.kansai-u.ac.jp
[2] Osaka Prefecture University, 1-1 Gakuen-cho, Nakaku, Sakai-shi, Osaka, Japan
morita@eco.osakafu-u.ac.jp

Abstract. Collaborative filtering is used for the prediction of user preferences in recommender systems, such as for recommending movies, music, or articles. This method has a good effect on a company's business. E-commerce companies such as Amazon and Netflix have successfully used recommender systems to increase sales and improve customer loyalty. However, these systems generally require ratings for the movies, music, etc. It is usually difficult or expensive to obtain such ratings data comparison with transaction data. Therefore, we need a high quality recommender system that uses only historical purchasing data without ratings. This paper discusses the effectiveness of a graph-partitioning method based recommender system. In numerical computational experiments, we applied our method to the purchasing data for CDs, and compared our results with those obtained by a traditional method. This showed that our method is more practical for business.

Keywords: Recommender system, Graph-partitioning, Collaborative filtering, Binary Data, Data mining.

1 Introduction

Recommender systems have become more important in terms of e-business as people access numerous e-commerce Web sites. These systems are used to present for items that users prefer. Recommender systems can be categorized into two main fields: collaborative filtering (CF) and content-based recommendation (CB). CF algorithms find similar users and decide on recommendation items using the ratings of those similar users. On the other hand, CB algorithms select recommendation items based on a match between content attributes and user preferences.

According to [1], one advantage of CF algorithms is that they can perform in domains where items have little available content associated with them. Another advantage is that CF systems can sometimes provide serendipitous recommendations. On the other hand, CF systems contain two fundamental problems: the user-item rating matrix is generally very sparse, and the first-rater problem is generated when users come across new or obscure items.

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 259–269, 2009.

The points highlighted in the above studies are understandable, and the existing CF algorithms, such as those that make recommendations on the basis of ratings data, show some good results. However, we need to point out some additional problems from the viewpoint of practical usage. First, CF approaches require a large amount of ratings data. Although firms have general POS data or panel data, it is unusual to have ratings data for items and obtaining such data from users leads to the incurrence of additional cost. From the viewpoint of practical usage, it is important to recommend something unexperienced and valuable to a user, by using his/her historical purchasing data. We think that it is more practical to develop an algorithm that uses the relationships between purchased (or used) items, although additional ratings data can also be used. Hence, we need a high quality recommender system that uses only binary data, such as historical purchasing data, without ratings. In this study, we propose a recommendation algorithm based on the graph-partitioning method by using the relationships between items. We also propose a method for selecting recommendation items we consider the extent of a user's preference based on the types of purchased items. We used the historical purchasing data for CDs in computational experiments [1].

The remainder of this paper is organized as follows. Section 2 reviews related works. Section 3 explains the recommender system using a graph-partitioning problem and recommendation strategies. Section 4 provides a brief overview of the traditional user-based and item-based CF approaches. Finally, Section 5 proposes an evaluation metric and reports the experimental results.

2 Related Works

Currently, the use of CF algorithms is the most popular approach in recommender systems. Over the last decade, many studies have proposed CF algorithms [2], [3], [4], [5], [6], and [7]. These algorithms can be broadly classified as user-based and item-based CF algorithms. The former are used to find users with similar characteristics, and the latter to find similar items from a data set. Both algorithms require considerable time to handle large online data sets. As an alternative, Rashid et al. proposed a two-step method that used k-means clustering and a CF algorithm, called ClustKNN [7].

In the first step of their method, users are divided into k similar groups using the k-means clustering. Then, the CF algorithm makes a recommendation using k surrogate users. The division of the search process into offline and online parts is a very intelligent means of cost reduction since the main problem is caused by the cost of online part.

Related studies on binary data have also been conducted by [8] and [9]. These are similar and propose scoring methods. In these methods, a predictive value is calculated on the basis of the weighted sum of the "votes" of other similar users.

[1] The data was obtained from the 2006 Data Analysis Competition, which was sponsored by the Joint Association Study Group of Management Science (JASMAC).

Although the Jaccard coefficient is used as a weight in [9], its framework is very similar to general CF approaches.

These CF approaches are very popular and have shown some successful results, but their problem is to require a large amount of rating data. Firms that record only panel data would want to have such a recommender system, so it is desirable to recommend something based only on the historical purchasing data for items.

This paper proposes a recommendation algorithm. Our algorithm is similar to ClustKNN in that both are two-step algorithms; however, our algorithm uses the graph-partitioning method as the first step to divide similar items into k groups. Note that ClustKNN divides users into k groups using k-means clustering. Data on the relationships between items are very sparse in general; hence, by using the graph-partitioning technique, we can obtain some meaningful subitem sets. We propose a new algorithm by using the subsets as the second step, whereas ClustKNN uses a general CF algorithm with k surrogate users.

3 Recommender System Using Graph-Partitioning Method

The recommender system domain comprises a user $u(u = 1, ..., n)$ and an item $i(i = 1, ..., m)$. Additionally, each user's preferences on item, $R_{u,i}$ can be explicitly given for the item i by the user u as 0 or 1. In this case, we suppose that 1 denotes purchase and 0 denotes no purchase. The system performs two types of computational tasks: the offline task of building a model using a graph-partitioning method and the online task of generating predictions and recommendations. In the offline task, the purchasing data are transformed into graphical data to express the relationships between the items. Subsequently, the graphical data is partitioned into subgraphs using a graph-partitioning method, in order to partition the items into several groups. In the online task, recommended items that have not yet been rated are decided for a target user. Using the relationships between the items obtained from the partitioned subgraphs improves the quality of the recommendations.

3.1 Application of the Graph-Partitioning Problem

Graph-partitioning problems are NP hard problems that require the partitioning of the vertices of a graph into several groups, in a way that maintains the balance of the vertical size and minimizes the sum of cutting costs of the edges between the groups [10]. The graph-partitioning problem is widely applied in several fields, including VLSI design, network design, and data mining.

In order to apply the graph-partitioning problem, we initially transform the purchasing data into a weighted undirected graph. Given a graph whose vertices correspond to the items, and the edges correspond to the relationships between the items purchased by the user. Therefore, it is desirable to give the weights which represent the strengths of the relationships between the items.

In this study, the weighted undirected graph G is defined as $G(V, E)$ where V is a set of vertices and E is a set of edges connecting two vertices in V.

Given a positive integer k ($k \leq m$), V is partitioned into k subgraphs—that is, $V = V_1 + V_2 + ... + V_k$ (disjoint union)—which maintains the balance of the vertical size, as shown in equation (1).

$$w(V_q) \leq \rho \cdot \frac{w(V)}{k}, \tag{1}$$

where the value of ρ is 1.03 from a parameter of METIS [11], and $w(V_q)$ and $w(V)$ denote the sums of weights of the vertices in V_q and V, respectively. Additionally, the main purpose is to minimize the sum of the costs of the edges between the subgraphs, whose cost is expressed as shown in equation (2).

$$\sum_{(x,y)\in E, x\in V_q, y\in V_p, p\neq q} c((x,y)), \tag{2}$$

where $c((x,y))$ denotes the cost of the edge between x and y.

Next, we determine suitable values for the costs and weights. The costs of the edges are determined by the strengths of the relationships between the items, such as "support," "confidence," [12], or the "Jaccard index" [13]. We used the support value to determine the costs of the edges, because the confidence value requires the directions of the edge. Moreover, the confidence and Jaccard index tend to increase if the number of users who rate the items, x and y, is few. In particular, if only one user rates two items, x and y, and no other user rates two items, the confidence and Jaccard index become 100%. This value widely influences the divided subgraphs. Therefore, we use the support value as the weights of the edges, $c((x,y))$, which is defined by

$$c((x,y)) = \frac{\text{the number of users purchasing items of both } x \text{ and } y}{\text{the total number of users}}. \tag{3}$$

Note that in actual experiments, the numerator of the above equation is used for cost determination, because the graph-partitioning problem can handle only positive integers in such cases. We give a unit weight for the weight of the vertex, because it is desirable that the number of items is the same in our method.

By applying the graph-partitioning method, the obtained subgraphs exhibit interesting properties. One of these properties is that the total weight of the vertices is almost the same at each subgraph because of the balance of the vertical size. In other words, every subgraph has almost the same number of items. Therefore, this property shows that we can recommend items in the same manner even if we select a subgraph randomly. Another property is that the relationships between the items contained in the same subgraph are stronger than those between items that belong to different subgraphs. In other words, the remaining edges show strong relationships compared to the eliminated edges. Therefore, the calculation costs can be reduced when recommending items because we can search only for items connected by edges.

3.2 Recommendation Strategies

Typically, the recommendation is represented by a ranked list of items that the target user will like. This type of recommendation is also known as a Top-N

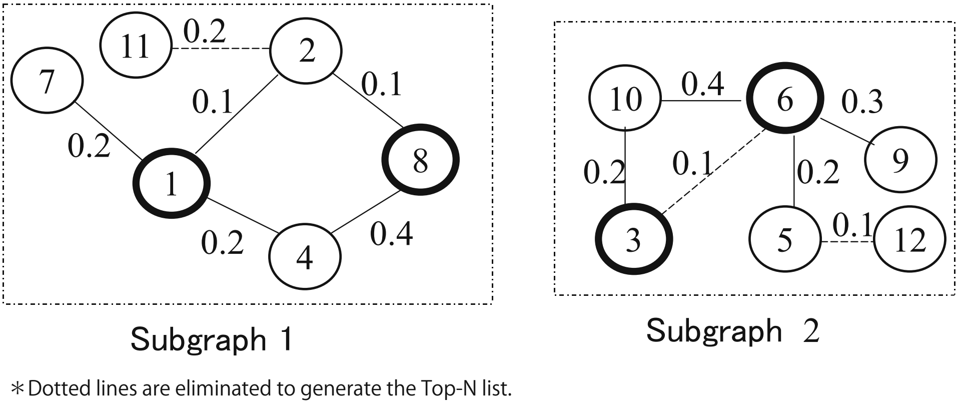


Fig. 1. Generating the Top-N list

recommendation [14]. The Top-N list is generated by sorting the results of the prediction rating on all the items not yet rated. In order to generate the Top-N list in our recommender system, we sort the edge cost totals.

We can illustrate the method of selecting the items, which consists of a Top-N list, by using the subgraphs. Figure 1 shows the subgraphs obtained from the graph-partitioning problems. The bold circles represent the purchased items, $z = 1, 3, 6, 8$. The edges between the purchased items z and non-purchased items that connect with z directly, $s = 2, 4, 5, 7, 9, 10$, remain, while the other edges are eliminated. Finally, the value of the recommendation is calculated as the sum of the costs of the edges that connect each s; for example, the value of the recommendation for $s = 4$ is 0.6, because from vertex 1 and vertex 8, two cost values, 0.2 and 0.4, are connected and added. The predicted value of each s is calculated in this manner.

When we generate the Top-N list, we adopt three types of strategies as follows.

Strategy 1. The Top-N list is generated by sorting the recommendation values for all of the candidate items.

Strategy 2. The Top-N list is generated using the D'Hondt method, which has been used for allocating seats in proportional representation systems.

Strategy 3. The Top-N list is generated by considering the subgraphs that belong to many non-purchased items.

Figure 2 shows a position of the three strategies. Strategy 1 emphasizes a stronger relationship between the recommended items and the purchased items. If this value is high, it shows that the item is strongly connected to the purchased items. However, if almost all of the purchased items have strong relationships with items that belong to the same subgraph, this strategy may select recommended items exclusively from the same subgraph. Strategy 2 decides on the number of recommendation items using the D'Hondt method applied to the number of subgraphs that belong to the purchased item. By using Strategy 2, we can determine the number of recommended items from each subgraph

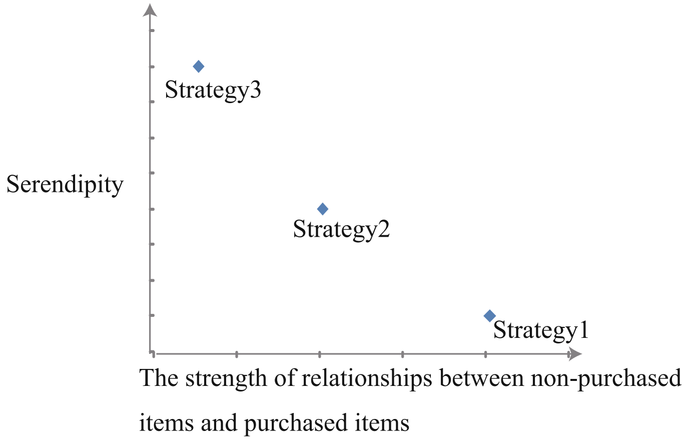


Fig. 2. The position of the three strategies

using the relative size, that is, the number of purchased items in each subgraph. Strategy 3 emphasizes serendipity. The recommended items are selected from subgraphs that barely belong to purchased items using the inverse of the ratio of the subgraphs that contain the purchased items. Later, we will compare these three types of strategies.

4 Recommender Systems Using Collaborative Filtering

In order to compare our proposed method with traditional CF systems, we describe two types of CF systems: user-based CF and item-based CF. We also describe a method that just uses random sampling and is the simplest available technique. We will now provide a brief overview of each algorithm.

4.1 User-Based CF Algorithm

In this algorithm, the rating prediction is obtained from the similarity of user preferences. The basic idea here is that users with similar characteristics will have similar preferences. This algorithm comprises a two-step process. First, the similarities between users are computed from the common items that have been purchased by both the target user and all the other users. Although several techniques are available to determine this similarity, we selected the Pearson correlation, which is the most commonly used technique [5]. The similarities of the users are computed by equation (4).

$$simU(u,tu) = \frac{\sum_{i\in I}(R_{u,i}-\bar{R}_u)(R_{tu,i}-\bar{R}_{tu})}{\sqrt{\sum_{i\in I}\left(R_{u,i}-\bar{R}_u\right)^2}\sqrt{\sum_{i\in I}\left(R_{tu,i}-\bar{R}_{tu}\right)^2}}, \quad (4)$$

where tu denotes the target user, and I is the set of items purchased by both u and tu. $\bar{R}_u$ is the average for a user u on all the items.

Next, the rating prediction $\hat{R}_{tu,ti}$ is computed by using the similarity among the users obtained from step one and determining a weighted average for the deviations from the selected user means:

$$\hat{R}_{tu,ti} = \bar{R}_{tu} + \frac{\sum_{u \in UC} (R_{u,ti} - \bar{R}_u) \cdot simU(u, tu)}{\sum_{u \in UC} simU(u, tu)}, \tag{5}$$

where ti denotes the target item, and UC is the set of users where $simU(u, tu) \geq \alpha$. Here an appropriate value is given to α from 0 to 1. The Top-N list is generated by sorting the results of the prediction rating.

4.2 Item-Based CF Algorithm

This algorithm also comprises a two-step process. It first computes the similarity between the items and then obtains a prediction by using this similarity. One fundamental difference between the similarity computations in the user-based CF and item-based CF is that in the former, the similarity is obtained from common items that are purchased by two different users, and in the latter, from users who have purchased two different items. In other words, the difference is either computed along the rows or the columns of the matrix. The similarity between two items i and b is computed by adjusting the cosine measure in equation (6), which was proposed by [6].

$$simI(i, b) = \frac{\sum_{u \in U} (R_{u,i} - \bar{R}_u)(R_{u,b} - \bar{R}_u)}{\sqrt{\sum_{u \in U} \left(R_{u,i} - \bar{R}_u\right)^2} \sqrt{\sum_{u \in U} \left(R_{u,b} - \bar{R}_u\right)^2}}, \tag{6}$$

where U denotes the set of users who have purchased both i and b.

Once the similarities between the items are computed, the rating space of the target user tu is examined to find all of the purchased items similar to the target item ti. The rating prediction $\hat{R}_{tu,ti}$ is computed using the weighted average by equation (7).

$$\hat{R}_{tu,ti} = \frac{\sum_{b \in N} (simI(ti, b) \cdot R_{tu,b})}{\sum_{b \in N} |simI(ti, b)|}, \tag{7}$$

where N denotes the set of items where $simI(ti, b) > \beta$. As with α, an appropriate value is given to β from 0 to 1. Typically, a threshold for the β similar items is used rather than that of all the items. The Top-N list is generated by sorting in the same manner as the item-based CF.

4.3 Methods of Random Sampling

This method of using only a random sampling for generating the Top-N list is simpler than the others. In experiments, this method is probably considered to be the standard of the lowest performance.

5 Experimental Evaluation

5.1 Evaluation Metrics

The F1-measure was used in our experiments. This is widely used in recommender systems and is calculated on the basis of *precision* and *recall*. *Precision* and *recall* are also widely used to evaluate information retrieval systems, and can be defined as below.

$$Precision = \frac{|\text{purchased items} \cap \text{Top-}N \text{ list}|}{|\text{Top-}N \text{ list}|}, \tag{8}$$

$$Recall = \frac{|\text{purchased items} \cap \text{Top-}N \text{ list}|}{|\text{purchased items}|}, \tag{9}$$

where $| \cdot |$ denotes the number of elements in the set and "purchased items" means the items that are involved in test sets.

$$F1 = \frac{2 * precision * recall}{precision + recall} \tag{10}$$

5.2 Data Set

In order to evaluate these methods, we used historical purchasing data for CDs. This data was sourced from a specific CD store in Japan. The collected data comprised data for two years and contained 1,200,000 records, 330,000 customers, and 50,000 items. We divided the purchasing data into the first and second years for use as training sets and test sets, respectively. We selected 1,340 users who purchased more than fifteen CDs in the first year and more than five CDs in the second year. We determined the target users, who consisted of 10% of the 1,340 users, by using a ten-fold cross validation. Therefore, the recommended items for the target users were determined by using the training sets.

5.3 Results and Discussion

We compared the proposed method with the other methods that were reviewed in the previous section by using the abovementioned evaluation metrics. For the experiments, we set parameters for the graph-partitioning and item-based CF methods. In the graph-partitioning method, the number of partitions was $k = 30$. This value was selected based on the results of preliminary experiments. In the cases of the user-based CF and item-based CF, we used $\alpha = 0.1$ and $\beta = 0.2$, respectively. These conditions were also determined by conducting preliminary experiments.

Tables 1 and 2 show the Top-N recommendation results with $N = 5$ and 10, respectively, using *precision*, *recall*, and the $F1 - measure$. The Top-N list generated by Strategy 2 of our proposed method, that is, "graph based (Strategy 2)," showed the best quality of all the metrics in both the Top-5 and Top-10

Table 1. Results for the Top-5 recommendations using *precision*, *recall*, and $F1-measure$

	Precision	*Recall*	$F1$
Graph based (Strategy 2)	0.0121	0.0078	0.0094
Graph based (Strategy 1)	0.0087	0.0052	0.0065
User-based	0.0082	0.0051	0.0062
Graph based (Strategy 3)	0.0053	0.0037	0.0043
Item-based	0.0016	0.0013	0.0014
Random sampling	0.0009	0.0005	0.0006

Table 2. Results for the Top-10 recommendations using *precision*, *recall*, and $F1-measure$

	Precision	*Recall*	$F1$
Graph based (Strategy 2)	0.0099	0.0131	0.0112
Graph based (Strategy 1)	0.0073	0.0088	0.0078
User-based	0.0061	0.0077	0.0068
Graph based (Strategy 3)	0.0051	0.0066	0.0043
Item-based	0.0014	0.0023	0.0017
Random sampling	0.0007	0.0007	0.0007

recommendation results. The second best quality was obtained by Strategy 1, ”graph based (Strategy 1).” A comparison of our method with the traditional CF methods reveals that the user-based CF delivers better performance than the item-based CF and Strategy 3. Overall, although the values of all the metrics are poor, they are better than the values obtained by random sampling. Therefore we could show that it is very difficult to select an item that will be purchased in the near future.

We can make some important observations based on the experimental evaluation of the proposed methods. First, the proposed method provides better quality recommendations than the other methods. This shows that it is beneficial to use the relationships between items expressed by a graphical structure. Second, in terms of the recommendations made, our proposed method provides better quality recommendations for items based on both the Top-5 and Top-10 recommendation results. In particular, the second strategy, which created the recommendation list by using the D’Hondt Method, obtained the best quality. This shows that the important thing is to reflect a balance between the major items and minor items in the preferences of the target users. Although the experimental results confirmed that these selection methods are suitable, the quality of the recommendations made by them was still somewhat low. Therefore, some improvements are needed to consistently generate a high quality Top-N list.

6 Conclusions and Remarks

In this paper, we proposed a recommender system that uses the graph-partitioning method and recommendation strategies that employ binary data without ratings. In the computational experiments, we illustrated that the proposed method provides better quality recommendations than the other methods. Our method provides better quality recommendations by using the relationships between items. However, we need to produce better quality recommendations in actual business applications.

It was an interesting challenge to use the D'Hondt method to select subgraph. It outperforms traditional methods, because of the selecting subgraphs method. It depends upon whether a customer likes major items or minor items. Therefore, it may be effective to incorporate some degree of customization. For instance, major items might be primarily recommended to some customers, while minor items are recommended to other customers. Then it would be necessary to identify customer groups to apply a more suitable strategy.

Another idea to improve the performance involves selecting items in the subgraph. They could be selected based on a strong relationship with existing purchased items. It might be better to select other items for specific customers.

We would like to incorporate these ideas to improve the recommendation performance in future works.

References

1. Melville, P., Mooney, R.J., Nagarajan, R.: Content-boosted collaborative filtering for improved recommendations. In: Eighteenth national conference on Artificial intelligence, pp. 187–192 (2002)
2. Hofmann, T.: "Probabilistic latent semantic analysis,'. In: Proceedings of Uncertainty in Artificial Intelligence, Stockholm (1999)
3. Pennock, D.M., Horvitz, E., Lawrence, S., Giles, C.L.: Collaborative filtering by personality diagnosis: A hybrid memory and model-based approach. In: Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence, pp. 473–480 (2000)
4. Herlocker, J.L., Konstan, J.A., Borchers, A., Riedl, J.: An algorithmic framework for performing collaborative filtering. In: Proceedings of the 1999 Conference on Research and Development in Information Retrieval (1999)
5. Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., Riedl, J.: GroupLens: An open architecture for collaborative filtering of netnews. In: Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work, pp. 175–186 (1994)
6. Sarwar, B., Karypis, G., Konstan, J., Reidl, J.: Item-based collaborative filtering recommendation algorithms. In: Proceedings of the 10th international conference on World Wide Web, pp. 285–295 (2001)
7. Rashid, A.M., Lam, S.K., Karypis, G., Riedl, J.: ClustKNN: A Highly Scalable Hybrid Model- & Memory-Based CF Algorithm. In: WEBKDD 2006 (2006)
8. Weiss, S.M., Indurkhya, N.: Lightweight collaborative filtering method for binary-encoded data. In: Siebes, A., De Raedt, L. (eds.) PKDD 2001. LNCS (LNAI), vol. 2168, p. 484. Springer, Heidelberg (2001)
9. Mild, A., Reutterer, T.: An improved collaborative filtering approach for predicting cross-category purchases based on binary market basket data. Journal of Retailing and Consumer Services 10(3), 123–133 (2003)
10. Garey, M.R., Johnson, D.S., Stockmeyer, L.: Some simplified NP-complete problems. In: Proceedings of the sixth annual ACM symposium on Theory of computing, pp. 47–63 (1974)

11. Karypis, G., Kumar, V.: Multilevel k-way Partitioning Scheme for Irregular Graphs. Journal of Parallel and Distributed Computing 48(1), 96–129 (1998)
12. Tan, P., Kumar, V., Srivastava, J.: Selecting the right interestingness measure for association patterns. In: Proceedings of the Eight ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 32–41 (2002)
13. Jain, A., Dubes, K.R.C.: Algorithms for clustering data. Prentice-Hall, Englewood Cliffs (1998)
14. Herlocker, J.L., Konstan, J.A., Terveen, L.G., Riedl, J.: Evaluating collaborative filtering recommender systems. ACM Transactions on Information Systems (TOIS) 22(1), 5–53 (2004)

Optimization of Budget Allocation for TV Advertising

Kohei Ichikawa[1], Katsutoshi Yada[1], Namiko Nakachi[1], and Takashi Washio[2]

[1] Kansai University, 3-3-35 Yamate-cho, Suita-shi, Osaka 564-8680, Japan
{ichikawa@rcss,yada@ipcku,da60437@ipcku}.kansai-u.ac.jp
[2] Osaka University, 8-1 Mihogaoka, Ibaraki-shi, Osaka 567-0047, Japan
washio@ar.sanken.osaka-u.ac.jp

Abstract. This research aims to present an analysis to optimally allocate advertising budgets based on single source data on consumers' views of TV advertising. A model of consumer behavior and an optimality criterion for the advertising budget allocation are proposed together with a GA based optimization algorithm. Through the analysis, we discovered some knowledge to improve the effectiveness of advertising for several products.

Keywords: advertising, optimal budget allocation, GA.

1 Introduction

Along the development of information technology in recent years, new media have appeared on Internet and mobile phones services, and these increase the variety of advertising media. Since the financial crisis began, many companies are under pressure to reduce expenses and earnestly striving to cut advertising costs. Under this situation, the cost effectiveness of diverse advertising media is becoming much more important. In particular, because TV advertising comprises the largest share of company advertising costs, maximizing cost effectiveness of TV advertising is becoming a top priority in many companies.

Marketing researchers throughout the world have done much research on TV advertising. Some of the earliest research was to develop an advertising effect model called MEDIAC [1]. MEDIAC is a model on the responses of individual consumers to advertising, and it is used to support decision making of advertisers. After this work, many advertising effect models were proposed in the marketing science field [2]. On the other hand, some researches on the synergy effect of cross media [3,4] and the time-series effect of advertising [5] have been conducted. However, almost all of these researches used data of advertising GRP and sales value, but made insufficient use of detailed panel data, and there is very little research on optimal budget allocation based on advertising effect models.

This research aims to an analysis to optimally allocate advertising budgets based on single source data on consumers' views of TV advertising. We propose the following analysis framework in order to support decision making on the optimal allocation of advertising budgets. First, in order to ease the advertising

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 270–277, 2009.

planning which includes a task to choose appropriate TV programs from the vast candidates for broadcasting of the advertising, we propose to categorize TV programs by cost for advertising and time slot when it was broadcasted. Next, we propose a novel model to estimate the number of purchasers on an objective product based on the single source data with the new TV program categorization, and derive the optimal resource allocation by using the model and genetic algorithm based optimization under the constraint of a given total advertising budget.

The rest of this paper is organized as follows. Section 2 describes the objective single source data and the new TV program categorization in this paper. Section 3 proposes our novel model of consumer's purchase and our new approach to derive the optimal allocation of advertising budget. Section 4 represents the results of the optimal allocation. Finally, the paper is concluded in Sect. 5.

2 Objective Data and New TV Program Categorization

This research uses single source data provided by Marketing Analysis Contest 2008 in Japan. This data contains information on 3000 individual consumers from February to March 2008. It includes each TV program viewed by every consumer during that period (5 weeks), records of all TV programs containing a TV advertising of each product, consumers' purchasing behavior for each product, their lifestyles, and so on. This data set enables to estimate impacts of the contact between an advertising and a consumer on its purchasing behavior.

Table 1. Each consumer's views of every TV program category, its purchase behavior and number of advertising broadcasted in every TV program category

Sample ID	holiday-night(A)	weekday-night(A)	...	weekday-night(B+)	Purchase
1	5	8	...	4	Yes
2	1	1	...	2	Yes
3	2	4	...	8	No
.	.	.	...	.	.
m	.	.	...	.	.
# of advertising	20	32	...	20	

In conventional analyses, a categorization named "time rank" of TV program has been frequently used to represent the cost required to broadcast advertising. It consists of four classes (A, B, B+ and C) based on the cost reflecting audience rate of each TV program. Here, we propose novel categorization of TV program named "time slot". It indicates attributes of their broadcasting time, and is represented by a combination of a) weekday or holiday and b) daytime, night or midnight. Each TV program is categorized by the combination of these "time rank" and "time slot" in this paper.

The aforementioned data set is transformed into a specific format for the latter estimation of the impact of the contact between a consumer and an advertising on its purchasing behavior. The data set is divided into data subsets according to objective products, and each data subset of a product contains instances each of which is a relation consists of frequencies of TV program categories viewed by a consumer and the consumer's purchase behavior. The purchase behavior is labeled as "yes" if the consumer has not purchased the products before the objective period of the data acquisition, but purchased it later. Otherwise, it is labeled as "no". Each row in Table 1 is an example of this relation. For example, the first row indicates that the consumer labeled as ID 1 purchased an objective product while viewing the TV program categorized as holiday-night(A) 5 times and so on. In the mean time, the number of the product's advertising broadcasted in each TV program category is counted for the latter use as indicated at the bottom row of Table 1.

3 Optimal Allocation of Advertising Budget

In this study, we address an optimization problem to find an advertising budget allocation pattern that maximizes a utility of advertising under a given budget based on a novel model to estimate number of purchasers. First, we build the model by using the data produced in the former section. Next, we define the utility to optimize the advertising budget allocation based on the model, and present an algorithm for the optimization.

3.1 Model to Estimate Number of Purchasers

To estimate the number of the purchasers of an objective product, we assume that a consumer, who actually purchased the product, reaches the purchase after viewing the advertising of the product at least three times. This assumption is based on three exposures theory [6]. In addition, we also assume that the chance of viewing advertising in a TV program category linearly increases depending on the number of advertising placed in every TV program category.

Let T be a set of all TV categories, and let I_t be the number of the advertising on a product placed in a TV program category t ($t \in T$). Let N be a set of purchasers of the product, and let V_{nt} be a frequency that each purchaser n ($n \in N$) viewed the advertising in the TV program category t with I_t. From the above assumptions, if the number of advertising in the category t is actually J_t, the number of purchasers E is estimated as:

$$E(J_t; t = 1, \ldots, T) = \sum_{n \in N} U\left(\sum_{t \in T} V_{nt}\frac{J_t}{I_t} - 3\right), \tag{1}$$

where the function U is a unit step function:

$$U(x) = \begin{cases} 0 & (x < 0) \\ 1 & (x \geq 0). \end{cases} \tag{2}$$

This model reflects the assumption that the expected number of viewing the objective advertising more than or equal to 3 contributes to the actual purchasing behavior.

3.2 Definition of Optimization Problem

By using this estimation E, we define an objective function to measure a utility consists of the advertising effectiveness and the cost-performance as follows:

$$F(J_t; t = 1, \dots, T) = E(J_t; t = 1, \dots, T) + \alpha P(J_t; t = 1, \dots, T). \tag{3}$$

Here, the number of purchasers E given by (1) is considered to represent the advertising effectiveness. P is cost-performance of the advertising given as follows, when the actual number of advertising in each category t is J_t $(t = 1, \dots, T)$.

$$P(J_t; t = 1, \dots, T) = \frac{E(J_t; t = 1, \dots, T)}{C(J_t; t = 1, \dots, T)} \tag{4}$$

where C is an advertising cost index arbitrary scaled by the audience rate of time rank when the advertising is broadcasted. The objective of the optimization problem is to find optimal J_t $(t = 1, \dots, T)$ which maximizes this utility F. α introduces a trade-off between the advertising effectiveness and the cost-performance. In this study, α is dynamically adjusted to equalize the contributions of both terms to the objective F according to the expected values of these terms during the optimization process.

3.3 Algorithm of Optimization

Assuming that the practically possible number of advertising to be placed in a TV program category is from 0 to 50 in each of 16 TV program categories, the number of possible advertising placements is 51^{16}, *i.e.*, around 2.1×10^{27}. Accordingly, the search space of our optimization problem is too large for the thorough search. In addition, the objective utility function is nonlinear and discrete, and the solutions of J_t $(t = 1, \dots, T)$ are constrained within sets of integers. Therefore, exact analytical methods, including linear programming and dynamic programming, are not applicable to this optimization problem.

Based on these observations, we employed genetic algorithm (GA) [7] which has wide and flexible applicability to the optimization problem. GA is one of meta-heuristic methods and suitable for the optimization problems having no known satisfactory problem-specific algorithms or better heuristic methods. Although GA does not always ensure the optimality of its solution, it is known to produce an acceptable suboptimal solution in most problems.

In GA, feasible candidate solutions are encoded in form of chromosomes, and it searches a suboptimal solution through crossover, mutation and natural selection of the chromosomes. The approach of GA is characterized by the following factors: encoding method for chromosome, crossover operation, mutation operation and selection operation. We encoded a placing pattern of advertising into a chromosome which is a T dimensional vector of integer numbers. The t-th element of the vector indicates the number of advertising placed in the TV program category t. For the crossover, a blend crossover operation [8] was adopted, because our chromosomes are expressed with integer number. Its offspring is generated by calculating values intermediate between two parent chromosomes. The crossover rate is set to be 80% in this study. We introduced two types of mutation operations. One is to find a completely new placing pattern. Through this mutation, the number of placing in a TV program category is completely changed. The second type is to refine a current pattern where the number of placing in a TV program category is slightly changed. The mutation rate of each type is set to be 1%. For the selection, we applied both elite selection and roulette selection. We pick up the top 20 chromosomes having larger values of the utility F by elite selection, and apply the roulette selection to the rest of the chromosomes in every generation. The roulette selection was employed to pick up candidate chromosome sets for the crossover and the mutation. The application of these operations determines new chromosome sets of the next generation.

4 Evaluation and Demonstration

4.1 Example Result of Optimization

Figure 1 and 2 are the original placing pattern $(I_t; t = 1, \ldots, T)$ and the optimized placing pattern $(J_t; t = 1, \ldots, T)$ for *Soyjoy* which is a healthy snack food. They show that the optimal placing pattern concentrates into particular categories such as weekday-midnight (C), holiday-morning (B) and holiday-night (C), while the original pattern covers a wide range of TV program categories. This means that the original pattern still has much room for improvement of cost-performance.

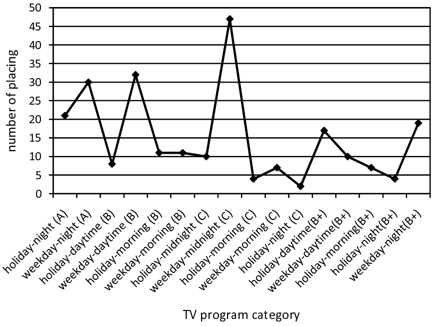


Fig. 1. Original placing pattern of Soyjoy

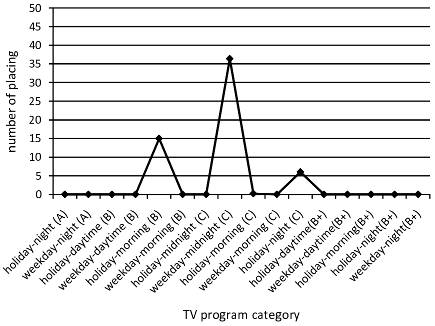


Fig. 2. Optimal placing pattern of Soyjoy

Table 2. Performance of original and optimal patterns for Soyjoy

		Original pattern (I_t)	Optimal pattern (J_t)
Utility of advertising	F	161.6	171.3
Number of purchasers	E	140	78
Cost-performance	P	0.168	0.724
Cost for advertising	C	833.69	108.62

Table 2 shows the comparison between the original and optimal patterns on the utility of advertising, the number of purchasers, the cost-performance and the cost for advertising appeared in (3) and (4). The number of purchasers of the original pattern is the actual number of purchasers observed in our data, while that of the optimal pattern indicates the estimated number E by (1). For both the original and the optimized patterns, the utility of advertising F is computed by (3) with $\alpha = 128.9$ which was determined through the optimization process. Table 2 indicates that the optimal pattern maximizes the utility of advertising F as the result of the improvement of cost-performance P with smaller budget C. The cost for advertising was cut down by 87.0% while the number of purchasers was reduced by 44.3%. Consequently, this resulted in 431 % improvement of the cost-performance. This optimal pattern suggests a good strategy to allocate the advertising budget efficiently to the TV program categories.

4.2 Comparison among Different Products

Comparison of the optimal patterns among different products suggests important characteristics of each product. This suggestion can be used to design the optimal allocation of advertising budget on each type of the products.

Figure 3 and 4 represent the optimal placing patterns of *Kuro-oolong-Cha* and *Nodogoshi-Nama* respectively. *Kuro-oolong-Cha* is healthy oolong tea similar to *Soyjoy*. In fact, those two optimal placing patterns, *Soyjoy* (Fig. 2) and

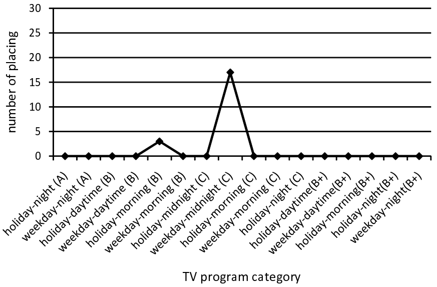


Fig. 3. Optimal placing pattern of Kuro-oolong-Cha

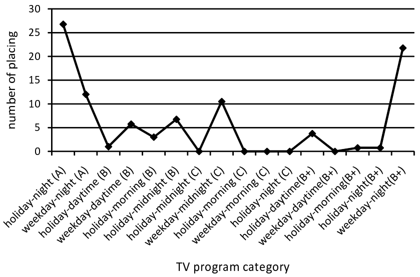


Fig. 4. Optimal placing pattern of Nodogoshi-Nama

Kuro-oolong-Cha (Fig. 3), are similar. On the other hand, *Nodogoshi-Nama*, which is a beer, has a completely different optimal pattern from the previous two patterns. The optimal pattern depicted in Fig. 4 has particular placement in holiday-night (A) and weekday-night (B+) together with the placements over a wide range of TV program categories.

This difference of the optimal patterns between these types of the products is considered to come from the difference of their target consumers. The insights suggested through this analysis can be used to design the optimal allocation of the advertising budget based on the characteristics of the objective products.

4.3 Optimization for Carryover Effects

An important issue for the advertising budget allocation is to take into account carryover effect of the advertising. This effect is that partial contribution of the advertising to the consumers' purchasing behavior lasts degressively even after a time period when the advertising is broadcasted. The advertising placement over multiple weeks is presumed to be further optimized by including its contribution.

In this section, we take into account the carryover effect and try to figure out more efficient advertising placement. Figures 1 to 4 indicate that the optimal pattern for *Nodogoshi-Nama* has a complex advertising placement over various TV program categories comparing with the other two products. This fact suggests the possibility to further increase the utility of the advertising budget allocation for *Nodogoshi-Nama* by focusing the placement on less TV program categories to reduce the advertising cost, if we can introduce some extra effective measure for this optimization. Here, we introduce the advertising carryover effects as the extra measure and determine the optimal advertising placement over an advertising period consisting of 5 weeks. However, the number of parameters required to represent the placement over the 5 weeks is too large even for the GA-based optimization. To alleviate this difficulty, we focused on only 4 TV time ranks, A, B, B+, C, instead of 16 TV program categories and selected only the 9 time ranks, A_2, $B{+}_2$, A_3, $B{+}_3$, B_3, A_4, $B{+}_4$, B_4, C_4, which are known to have large contribution to the consumers' purchasing behavior based on our preliminary analysis. The suffixes of the time ranks represent week IDs over the 5 weeks.

Figure 5 and 6 show the original placing pattern in our target single source data and the optimal pattern of *Nodogoshi-Nama* over the 9 time ranks respectively. The original pattern has a major placement at A_2 and $B{+}_2$, while the optimal pattern has at A_4. In Japan, there are two typical advertising placement patterns. One is a placement at various TV program categories over a long term to appeal the objective product over widely targeted consumers, and the other is a placement at a particular category within a short term to appeal to focused consumers. As easily understood, the original placement of *Nodogoshi-Nama* adopts the former pattern and the above result suggests that *Nodogoshi-Nama* should be appealed to focused consumers to enhance the utility of its advertising.

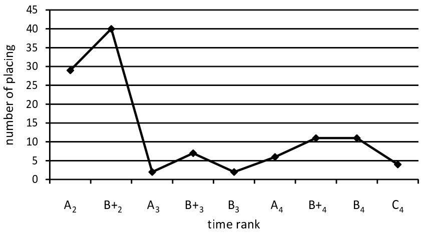


Fig. 5. Original time series placing pattern of Nodogoshi-Nama

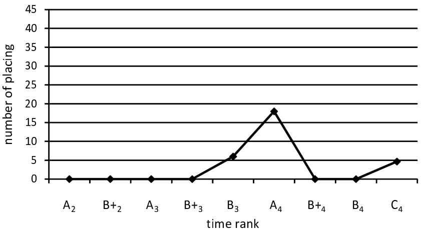


Fig. 6. Optimal time series placing pattern of Nodogoshi-Nama

5 Conclusion

In this paper, we investigated a new framework for the optimal allocation of advertising budgets. The framework we presented in this paper clarifies that the optimization of the advertising placement pattern according to a utility measure of advertising dramatically improves the cost performance of the advertising for various products. However, many issues remain. This framework does not consider some other crucial effects of advertising such as reach, recognition and cross media effects. There is a need to develop a further integrated framework in the future, which includes various factors.

Acknowledgments. This work is supported by DSI Program at Kansai University and Grant-in-Aid for Scientific Research on Priority Areas (21013032).

References

1. Little, J., Lodish, L.: A media planning calculus. Operations Research (17), 1–35 (1969)
2. Rao, A.G., Miller, P.B.: Advertising/sales response functions. Journal of Advertising Research 15, 7–15 (1975)
3. Gatignon, H., Hanssens, D.: Modeling marketing interactions with application to salesforce effectiveness. Journal of Marketing Research 24, 247–257 (1987)
4. Gopalakrishna, S., Chatterjee, R.: A communications response model for a mature industrial product: Application and implications. Journal of Marketing Research 29, 189–200 (1992)
5. Naik, P., Raman, K.: Understanding the impact of synergy in multimedia communications. Journal of Marketing Research 40, 375–388 (2003)
6. Krugman, H.: Why three exposures be enough. Journal of Advertising Research 12, 11–14 (1972)
7. Holland, J.H.: Adaptation in Natural and Artificial Systems: An Introductory Analysis With Applications to Biology, Control, and Artificial Intelligence. The MIT Press, Cambridge (1992)
8. Eshelman, L.J., Schaffer, J.D.: Real-coded genetic algorithms and interval-schemata. In: Whitley, D.L. (ed.) Foundation of Genetic Algorithms, vol. 2, pp. 187–202. Morgan Kaufmann, San Francisco (1993)

Cover All Query Diffusion Strategy over Unstructured Overlay Network

Yoshikatsu Fujita, Yasufumi Saruwatari, Masakazu Takahashi, and Kazuhiko Tsuda

Graduate School of Business Scineces, University of Tsukuba, Tokyo
3-29-1, Otsuka, Bunkyo-ku, Tokyo 112-0012 Japan
fujita@fw.ipsj.or.jp,
{saru,masakazu,tsuda}@gssm.otsuka.tsukuba.ac.jp
http://www.gssm.otsuka.tsukuba.ac.jp

Abstract. We have studied query diffusion strategy to cover all the nodes over unstructured overlay network. Although our previous work [1] covers 80% of nodes over the power-law network, we step further to minimize the number of left-behind nodes. In order to propagate messages to overall network, we assume the best case to choose the shortest path between every node pair. This is aimed for studying the optimal message path as a whole, reducing the query diffusion cost which is equal to the sum of minimum shortest path length. We have studied the characteristics of message propagation behavior, and that our proposed strategy can be applied for contents delivery over unstructured overlay network.

1 Introduction

In Japan, with rapidly growing demand for broadband contents delivery over the Internet, HD contents download service has been launched in the end of 2008 in Japan by acTVila corporation. This company has been established by five leading Japanese TV set manufacturers aiming for presenting TV-friendly contents which is easier to view, operate, pay and other transactions mainly on TV display. This is an epoch-making event in the Japanese Internet history because broadband stream (more than 10Mbps) delivery has become standard service even for TV viewers. This clearly shows that not only PCs but TV sets become more and more important devices as a broadband contents receiver over the Internet. Sooner ore later, millions of TV sets will access to the Internet searching for their favorite broadband contents, which causes unprecedented huge amount of traffic to the backbone network.

However, when it comes to the matter of how we should deliver broadband contents over the Internet, the total throughput will be determined by any bottleneck somewhere between contents provider to consumers. For example, even a user purchases 100Mbps optical fiber service, one's requested contents comes from distant server only 100Kbps because of narrow path somewhere over the Internet. Most of proxy servers distributed over the Internet are aimed for static contents like homepage objects, and are not tuned for broadband stream delivery.

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 278–284, 2009.

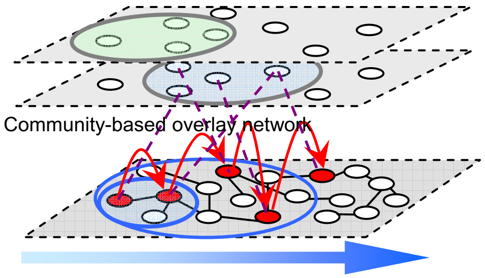


Fig. 1. Query Message Propagation on the Overlay Network

For instance, once many users try to pull large video streams at the same time, it is clear that backbone network is easily falls into overflow. This requires a new contents delivery technology which supports a huge simultaneous access transaction for broadband objects.

For this problem, we have proposed community-based overlay network over the Internet, and to manage generated traffic within that community, to be new message propagation method to cover the whole network [1]. This technique is based on the fact that any link status on the Internet follows power law distribution [2]. For example, one of the most popular pure P2P network Gnutella has been analyzed that its nodes' outgoing degree can be expressed as $P(k) \sim \mathrm{k}^{-\tau} (\tau \geq 0)$ [3]. However, such an unstructured network is not manageable in nature, and makes it difficult to apply for contents delivery for its fundamental network architecture. In our previous study, we have employed percolation theory [4] that is mainly studied in physics, to model the "percolating information" for query message delivery over pure P2P network. In other words, the query message released by client seeking for requested contents is not merely used for this purpose, but we have defined a new "reverse-query" message to find any client who needs a certain contents and try to apply the percolation theory to manage the generated traffic. This will lead to reduce the explosive P2P query traffic while maintaining fairly high clients cover rate over our proposed overlay network. An outlook of our model is shown in Fig.1.

2 Related Work

2.1 Contents Delivery Network

In order to deliver broadband contents over the Internet, CDN (Contents Delivery Network) architecture [5] and contents distribution algorithm for replication [6] are actively studied. But such CDN solutions for large scale contents delivery faces difficulty because the number of acceptable simultaneous access is almost determined by hardware specification of cache servers, and this falls into optimal cache server distribution problem with considering dynamic request load balance under the exact forecast of contents popularity and hardware availability. This is also regarded as a big issue for realizing broadcast type traffic over the Internet. Existing technique to handle telephone call over the telephone network is specific for point to point traffic, and it is not applicable for clearing simultaneous access to a contents sever, that makes it difficult to deliver broadband contents to many clients.

Recently, file sharing application over P2P network has been pervasive and it plays an important role as contents distribution infrastructure. When we apply P2P network for contents delivery, under the pure P2P architecture which has no center server, it is major concern how to find available resources and required contents from all over the network. In this field of study, such projects as CAN[7] , Chord [8] are trying to employ Distributed Hash Table (DHT)[9].

However, in order to apply for commercial services, there exist more problems in this DHT solution.

(1) When peers join/leave the network, they have to takeover management table to some other peers, results in heavy overhead.

(2) Network structure becomes complicated.

In addition, traffic generated by those P2P nodes has been increasing more and more, whose amount of load gives serious impact to today's ISP backbone network.

2.2 Reverse Query Diffusion

In our previous study [1], we proposed new network architecture for query message delivery which covers almost 80% nodes of overlay network, but still exist the left-behind nodes that never receive any query messages. Let us review the results of our idea to propagate a query message over the unstructured peer-to-peer network called “Reverse Query Mechanism” [10]. We have defined new “do-you-need” query and shown that contents can be delivered just relaying this query from server to peer to peer.

Macdonald et al. proposed to express the significance and network topology by MST (Minimum Spanning Tree) [11], but this has drawback that the message route depends on the value of weigh definition. Our study reveals that the value of minimum shortest path will not increase even without giving such weigh. We assume the best case to relay the query message by the shortest path between every node pair. This means that the amount of shortest path length does not increase, without any weighing operation.

3 Efficient Contents Delivery over Community-Network

When we regard community-network as an overlay network, the network structure will hold the nature of unstructured power-law overlay network. Our previous paper revealed that cover rate 80% to 90% had been achieved by our model, but it was difficult to cover all the nodes in the network. This is because we pass on the query without the knowledge of the whole network structure, and choose the query receiver just randomly. In other words, it is difficult to choose the best node which will be optimal for message propagation even we have the knowledge of whole network. This is true even for the central P2P architecture, because the ad-hoc behavior of every node is not predictable and we cannot rebuild the delivery path upon the leaving node activity. In order to present the solution for this problem, we assume the best case to choose the shortest path between every node pair. As we focus on the power-law nature of overlay network, we have studied how the video contents delivery should be deployed over this network, specifically paying attention on the required delivery time to cover all the nodes joining this overlay network.

3.1 Distribution of the Shortest Path and Bottleneck

First, we try to generate the overlay network which we employ as a contents delivery platform. Some parameters are shown below.

Data generation platform: GNU Octave 3.0

Graph Drawer: Pajek 1.4 [12]

Next, we examined the generation results. The distribution example of the shortest path length between each node pair is 2, is shown in the Fig. 2. This network consists of 1,000 nodes, 3,435 arcs and the probability of nodes' degree to be k is expressed as $P(k) = 0.49\mathrm{k}^{-1.51}$.

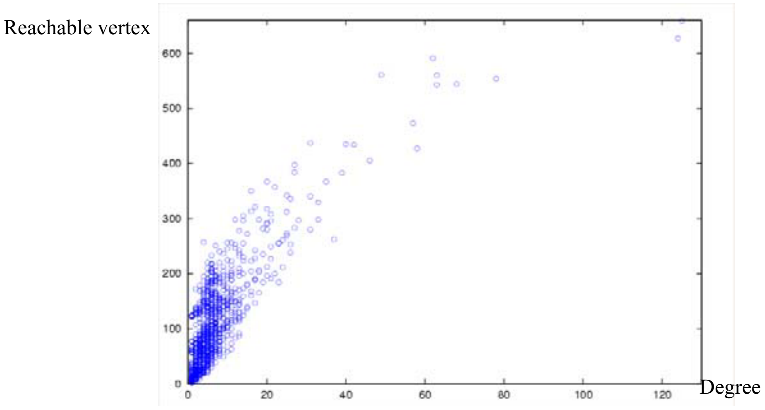


Fig. 2. Distribution of Shortest Path (SPL=2)

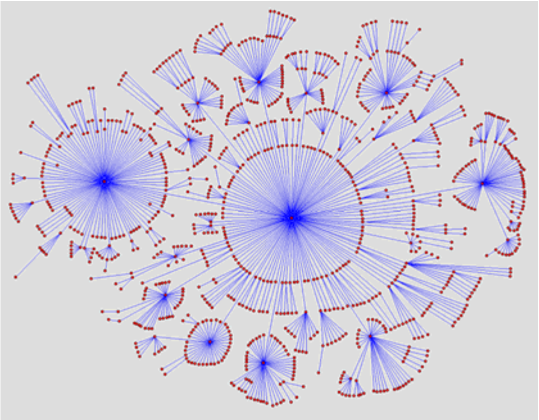


Fig. 3. Network by Shortest Path Branch

For improving the analysis, we assume that "Do-you-need" message propagation along the shortest path, because when we deliver the message, it is better to find the route that would not take detour. Then we assume the message will be relayed along with the shortest path and try to draw the graph how the message has been passed on from the origin. Our results are drawn in Fig 3.

As shown in this fig.3, the bottle-neck that would delay the message delivery is in the midst of nodes that are used by many nodes as a part of the shortest path. We can find that such a node will be an obstacle for message delivery because the relay probability is limited only for constant value. This giant node cannot pass on the message to every neighbor. Such a node possibly falls into not only drawing heavy accesses, but CPU overload for transaction. For instance, "hub" airport like New York boasts the concentration of airline flow. By employing our proposed strategy for "Do-you-need" message delivery, we can solve this bottleneck problem.

We have studied the average TTL to cover all the node is at most seven, to implement the shortest path delivery by Fig.3. As we have shown in the previous chapter, we can propagate the message to all the nodes in the overlay network.

3.2 Model Customization

In our previous study, when message is delivered on the shortest path, any nodes which behave as hub can be regarded as bottleneck. Therefore, we will employ algorithm to avoid such hubs for routing the message. When we look for the message path, we will force to avoid the top-three most accessed nodes in selecting the shortest path. This improves to lower the relay probability and minimize the message TTL value. The results for applying this algorithm is shown in Fig. 4

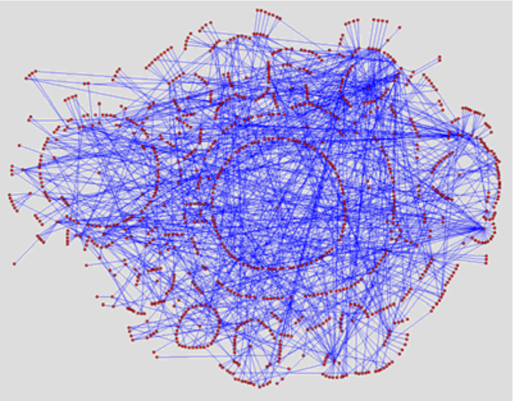


Fig. 4. Shortest Path Branch Avoiding Hubs (top-three)

We can reduce the usage frequency of hub by avoiding the most frequently used nodes from the message path as in Fig.5. In this case, the maximum value of shortest path length remains just eight, which means we can cover the whole network without increasing the delivery cost. While this proposal assumes the shortest path, we can

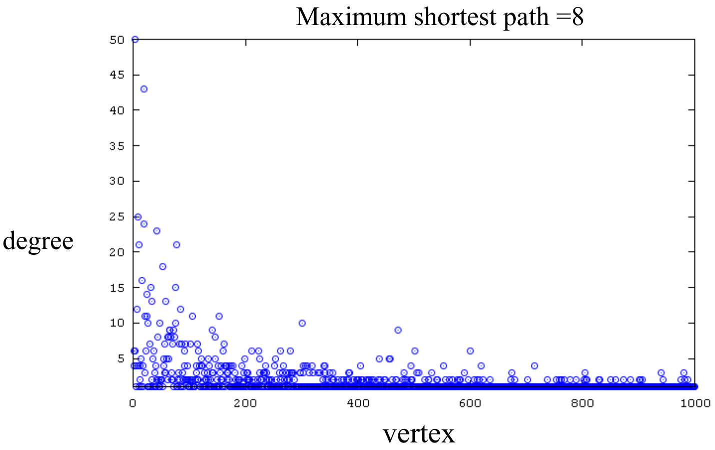


Fig. 5. Vertex Degree Distribution (Avoiding top-three hubs)

expect the similar effects when we apply the same algorithm in choosing the message path. This is to improve the message delivery overhead on our overlay network.

Next, we will consider applying this result for the community-based overlay network model. As already shown above, message delivery traffic is concentrated on the hub nodes when we try to propagate messages on power-law overlay network. And if we select the "best routing path," any message can reach every node on the network within maximum of shortest path length (hops) of eight.

In addition, when we regards the relay probability as a parameter, we got the combination of relay probability 0.3 and hop count 5 produces cover rate 80%, but it was difficult to percolate the message to all the nodes on the network. In order to overcome this limitation, our proposed method adds the condition that "preferentially choosing nodes other than hub" upon simple probabilistic decision. This leads to deliver messages to all the nodes while reducing the overall traffic. This algorithm can be implemented just to add an inquiry step whether its opponent nodes are hub or not, before sending message.

4 Conclusion

In this study, we have studied query diffusion strategy to cover all the nodes over unstructured overlay network of "Do-you-need" model, improving the cover rate. This is to propose how to propagate the query message to all over the network for video contents delivery. This shows that our proposed method can deliver message without increasing the maximum value of shortest path length. More specifically, a message can go through by getting around the most accessed hubs and maintaining the maximum value of shortest path length.

Also, our previous study shows that cover rate 80% can be achieved but difficult to cover all the nodes. We propose to apply our method for video contents delivery over the overlay network, that is, to find the distribution of the distance between every node pair and estimate the required time for contents delivery. This proves our proposed method can increase the efficiency of contents delivery over the overlay network. For commercial application example, we can attach video clip in the reverse-query message, which is possible to use this mechanism as electric flyer distribution.

References

[1] Fujita, Y., Saruwatari, Y., Yoshida, J., Tsuda, K.: Query message delivery over community-based overlay network. In: Apolloni, B., Howlett, R.J., Jain, L. (eds.) KES 2007, Part II. LNCS (LNAI), vol. 4693, pp. 1354–1361. Springer, Heidelberg (2007)

[2] Albert, R., Barabasi, A.L.: Statistical Mechanics of Complex Networks. Reviews of Modern Physics 74, 47–97 (2002)

[3] Faloutsos, M., Faloutsos, P., Faloutsos, C.: On Power-law Relationships of the Internet Topology. In: ACM SIGCOMM, pp. 251–262 (1999)

[4] Stauffer, D., Aharony, A.: Introduction to Percolation Theory. Taylor and Froncis, London (1994)

[5] Abrams, M., Standridge, C.R., Abdulla, G., Williams, S., Fox, E.A.: Caching Proxies: Limitations and Potentials. In: Proceedings of 4th International World Wide Web Conference, pp. 119–133 (1995)

[6] Li, Y., Liu, M.T.: Optimization of Performance Gain in Content Distribution Networks with Server Replicas. In: SAINT2003 Proceedings, pp. 182–189 (2003)

[7] Ratnasamy, S., Francis, P., Handley, M., Karp, R., Schenker, S.: A Scalable Content-addressable Network. In: Proceedings of the 2001 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications, pp. 161–172. ACM Press, New York (2001)

[8] Stoica, I., Morris, R., Karger, D., Kaashoek, M.F., Balakrishnan, H.: Chord: A Scalable Peer-to-Peer Lookup Service for Internet Applications. In: Proceedings of ACM SIGCOMM, pp. 149–160. ACM Press, New York (2001)

[9] Sunaga, H., Hoshiai, T., Kamei, S., Kimura, S.: Technical Trends in P2P-Based Communications. IEICE Trans. Commun. E87-B(10), 2831–2846 (2004)

[10] Fujita, Y., Yoshida, J., Tsuda, K.: Reverse-query mechanism for contents delivery management in distributed agent network. In: Khosla, R., Howlett, R.J., Jain, L.C. (eds.) KES 2005. LNCS (LNAI), vol. 3684, pp. 758–764. Springer, Heidelberg (2005)

[11] Macdonald, P.J., Almaas, E., Barabási, A.-L.: Minimum spanning trees on weighted scale-free networks. Europhys. Letters 72(2), 308–314 (2005)

[12] Pajek, `http://vlado.fmf.uni-lj.si/pub/networks/pajek/`

Extracting the Potential Sales Items from the Trend Leaders with the ID-POS Data

Masakazu Takahashi[1], Kazuhiko Tsuda[1], and Takao Terano[2]

[1] Graduate School of Business Sciences, University of Tsukuba
3-29-1 Otsuka, Bunkyo-ku, Tokyo 112-0012, Japan
{masakazu,tsuda}@gssm.otsuka.tsukuba.ac.jp
[2] Dept. Computational Intelligence and Systems Science, Tokyo Institute of Technology
4259 Nagatsuda-Cho, Midori-ku, Yokohama 226-8502, Japan
terano@dis.titech.ac.jp

Abstract. This paper, we focus on recommendation functions to extract the high potential sales items from the trend leaders' activities with the ID (Identification)-POS (Point-Of-Sales) data. Although the recommendation system is in common among the B2B or B2C businesses, the conventional recommendation engines provide the proper results; therefore, we need to improve the algorithms for the recommendation. We have defined the index of the trend leader with the criteria for the day and the sales number. Using with the results, we are able to make detailed decisions in the following three points: 1) to make appropriate recommendations to the other group member based on the transitions of the trend leaders' preferences; 2) to evaluate the effect of the recommendation with the trend leaders' preferences; and 3) to improve the retail management processes: prevention from the stock-out, sales promotion for early purchase effects and the increase of the numbers of sales.

Keywords: Recommendation Systems, Dual-Directed Recommendation, Collaborative Filtering System, Customer Preference, Trend Leader, Service Science and Management Engineering.

1 Introduction

There are various reasons for the store operation inefficiency in the retail business, especially small to mid size companies in Japan; one of the reasons for the issue is luck of the timing understanding for the items that customers needs. For instance, it comes from the stock out of the items for insufficient understandings. This issue will affect the whole retail industry not only to the whole seller but the manufacturers; therefore, we need to understand the both the preferences and the timings for the customers. As for the understanding for the preferences, the collaborative filtering system is one of the extracting algorithms for the preferences, and is adopted among the B2B or B2C industries [1].

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 285–292, 2009.

2 Related Works; The Conventional Recommendation Systems

Although the recommendation system is in common among the B2B or B2C businesses [2] [3] [4], the conventional recommendation engines provide the proper results, so that this issue should improve with new criteria when we apply to the retail industry. The customers of the retail industry are composed from with the diversity as following points; preferences, income, and number of family, therefore, we need to care generating the recommendation information not only with the properly but with the diversity.

2.1 Collaborative Filtering Systems

The collaborative filtering system has following characteristics of the algorithms; Easy to data convert, Affordability for the huge data, such as the data depend system. But, on the other hand, this system has a tendency influenced for the majority answer and both the items and the customers need to extract the adequate recommendation information as the negative points[5][6][7]. Therefore, a large amount of information related to consumers is required. From this large amount of the information, appropriate characteristics are extracted and the users are categorized into an appropriate segment, on this bases and appropriate product group needs to be identified for long tail business.

The collaborative filtering system is one of the techniques for providing suitable items from the customer's preference as the recommendation information among the many candidate items, and forecasts the customer's preferences from both the customer's activities and the purchase record. The feature of this method is to evaluate the customer's preference concerning item information and to generate the group of the customer preference with the similarity [1] [8].

Recommendation information is provides among items with the high evaluation from another customers but is not purchased yet with similar preferences. Therefore, the steps for generating the recommendation information is necessary as follows; (1) correct many purchase history for the customer, (2) Make the group and retrieve the customers who bought the same item, when a certain customer purchased the item, (3) Generate the recommendation information based on the item group among the same group customer. In this method, it doesn't analyze it concerning the content of contents at all. Therefore, it has the feature that the restriction concerning the recommendation object doesn't exist, and the situation in which only similar information to information evaluated in the past is recommended can be evaded by using other customers' evaluation information. That is, collaborative filtering system is a mechanism that customer's community is generated only from the purchase information without the analysis of contents of the item. Moreover, the problems of the collaborative filtering system are summarized as follows [7] [9];

- Both large number of customers and the amount of contents are required
- Same recommendation information is generated if the items to recommend are little.
- The items already bought only to be recommended.
- Impossible to recommend to those who have preferences different from the ordinary.

- Cannot prevented from influencing from mis-inputting
- New item is recommended until someone gives the evaluation even if the item was registered
- Can not connect the customers with the different ID even though they have the similar contents for without analyzing for the contents,

As for the forecasting for the new items sales number, Nakamura evaluated and classified the characteristics of the new items with the market reflection data such as the ID-POS data [10]. The conventional evaluation methods such as the trial repeat model only indicates the characteristics for the items that purchased repeatedly but, are insufficient for the customer classification or the recommendation from the ID-POS data classification [11].

3 Extracting the Trend Leaders

One of the conventional methods for understanding the timing for the customer needs is the time-series data analysis with the quantitative data such as the POS data. These methods are good for the inventory controlling such as the stable demand from the customer needs because of the sufficient time-series data and able to mark high score ratio for the demand forecasting. On the other hands, it is hard to forecast the demand such as the new items or the items that demand rapidly stood up because of the insufficient of the forecasting base data. To generate recommendation information with understanding both the customer preferences and the timing for the needs, it is required to not assuming the existence of the items; therefore, it comes to consider the existence for the new items. Moreover, it is expected to prevent the stock out to share the demand information among the retail industries; therefore, we referred "Dual-Recommendation" because of providing the recommendation information not only to consumers but also to manufacturers and retailers [12]. Therefore, this paper proposes the algorithm that detect to the timing of the item needs efficiently with the ID-POS data gathered from the local super market in Japan. Especially, to understand the timing of the item needs, we make use of the concept of the trend leader to forecast the demand of the new item. It is difficult to detect the ability with the trend leader of the item selection that foreseen the fashion in advance from the numerical data such as the ID-POS. Moreover, the candidates for the trend leader include the customer only with short span curiosity, so that an efficient selection method of detecting the trend leader is required.

3.1 Data Attributes and Demography

We make use of the ID-POS data to extract the trend leader. Table.1. indicates the data attributes gathered from the local retail store in Japan. This paper, we focus on the new registered items during the period to detect the items that will increase the number of sales.

Fig.1. indicates the relation between the number of Items for the Newly registered, the Purchased items, and the Increased items. This also indicates the survivability for the newly registered items. From the figure, about 50% of the newly registered items could not increase the sales number.

Table 1. Data Attributes

Term	2007/05/01~2007/12/31
Gathered Data	POS Transaction
	Reward Card Transaction
Category	Super Market
Number of Stores	3
Site	Shimane pref., Japan

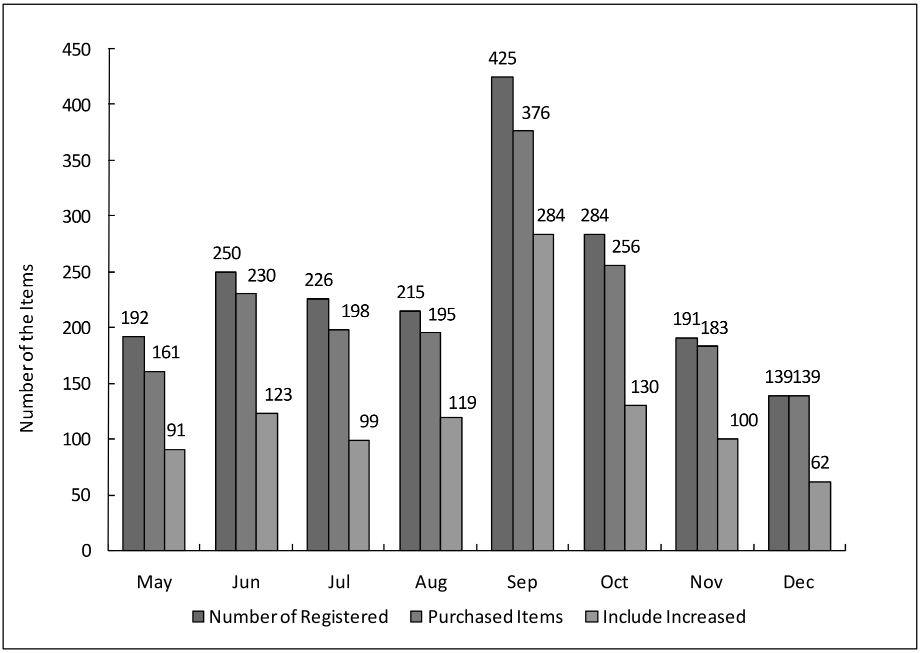


Fig. 1. Number of the Items for the Registered, the Purchased, and the Sales Increased

3.2 Extraction Methods

To understand both the customers' preference and the purchase timing at once, we have set up the dual-directed recommendation system that is able to consider both the new items and the stock outs with the Taste [13]. We have defined the trend leaders who bought the items at the initial stage were able to maximize the sales amount and defined the index of the trend leaders from those customers' activities and put those into the recommendation filters. This recommendation engine based on the functions for the proposed extracting method of the trend leaders is with the efficiency and with the diversity compared to the conventional recommendation information.

The conventional forecasting methods for the new item sales number have been based on the sales results of the similarity [14]. It is good for the new items that have fewer criterions both to avoid the stock out and to achieve the large scale sales with some indexes for forecasting of the maximize sales items from the initial demand. At the same time, this forecasting will good for the inventory controlling to avoid the stock out with the understanding the timing for the customer needs with sharing the item demand information among the retail industries; the whole sellers and the manufactures. Then, this paper focuses on extracting the items that can be maximized sales number with the index of the trend leader. One the other hand, it is difficult to detect the ability with the trend leader of the item selection that foreseen the fashion in advance from the numerical data such as the ID-POS.

To measure the indexes of the trend leader for each item, we take the following criteria;

- How first purchased items from the release
- How many purchased items that increasing of the sales number

Let N_i as the new item that the customer i purchases and M_i as the items that increased the sales number, the index of the trend leader (TL) is given as follows;

$$TL = \frac{\sum_{j \in M_i} \frac{1}{t_j}}{N_i} \qquad (M_i \in N_i) \tag{1}$$

3.3 Extraction Results

Table.2. indicates the results demography for the index of the trend leader (ITL) based on the extraction procedure. The maximum of the ITL was 6.91, and minimum ITL was 0 and the 8,707 customers in the 9,141customers became 1 or less for the index. The customer exceeded 6 or more was 3.

Table 2. Demography for the Index of Trend Leader

Index	Customer	%
~1	8,707	95.25%
~2	360	3.94%
~3	45	0.49%
~4	20	0.22%
~5	4	0.04%
~6	2	0.02%
~7	3	0.03%
total	9,141	

4 Recommendation Engines Comparison

The index of the trend leader is one of the reference values, so that it is necessary to set the threshold of the index properly according to characteristics of the region, items, customer attributes, and sales policy. Then, index of more than 2 is assumed to be a threshold, both to extract the trend leader and to recommend the items. In this paper, we took Taste [13] for the recommendation engine. As for the method for the similarity, the correlation coefficient was originally used, and changed the similarity method into the cosign distance from the correlation coefficient for holding many indexes of the similarity.

4.1 Evaluation for the Systems

Fig.2. indicates the results for the evaluation between the default recommendation engine and the improved one. The w/TL indicates the recommendation results based on the index of the trend leader in the figure.

This indicates the average increase ratio for the sales number at least 1.43% rise with the proposed algorithms. Table.3 indicates the recommendation results

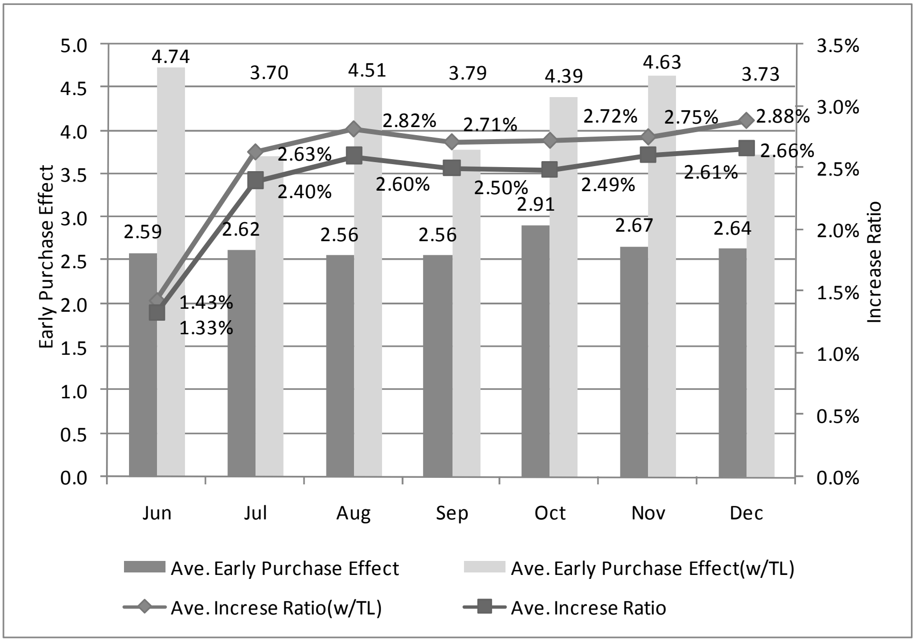


Fig. 2. Ave. Increasing Ratio for the Gross Purchase Number and the Early Purchase Effect

Table 3. Recommendation Results Comparison

	Max	Min
Average Increase Ratio(with/ITL)	2.88%	1.43%
Average Increase Ratio	2.66%	1.33%
Average Early Purchase Effect(with /ITL)	4.74	3.70
Average Early Purchase Effect	2.91	2.56

comparison matrix. As for the average early purchase effect, to recommend the items that the trend leader purchased are expected the distinct results for the item sales maximization.This early purchase effect is good for not only recommendation for the right time but ready for the stock out of the items.

From the results of the evaluation with the recommendation based on the trend leader, we have figure out the following issues; As the benefits for the recommendation with proposed indexes, detect the items that will increase sales number in the future,

- Prevention from the stock out
- Sales improvement for early purchase effect
- Increase of the sale number

As a result, we have succeeded to extract both the trend leaders from their purchase activities and the customer groups that became the trend leaders' candidates from the ID-POS data. We found out the items that will become the high potential for the increasing sales number in the future with the proposed indexes. The recommendation information from the trend leaders with high score made increase the sales number and shortened the days until purchasing.

5 Conclusion

In this paper, a basic research project in relation to the ID-POS transaction data analysis was described. Finding out the formula for extracting the trend leaders among the customers, that which confirm that there is a possibility to make appropriate recommendations to the other group member based on the transitions of the trend leaders' preferences.

From the results of the evaluation with the recommendation based on the trend leader, we have figure out the following issues;

- Prevention from the stock out
- Sales improvement for early purchase effect
- Increase of the sale number

As the result, we have succeeded to extract the trend leaders among the customers, that which confirm that there is a possibility to make appropriate recommendations to the other group member based on the transitions of the trend leaders' preferences and confirmed the effect of the recommendation with the trend leaders' preferences.

We are able to make detailed decisions in the following three points: 1) to make appropriate recommendations to the other group member based on the transitions of the trend leaders' preferences; 2) to evaluate the effect of the recommendation with the trend leaders' preferences; 3) to improve the retail management processes: prevention from the stock-out, sales promotion for early purchase effects and the increase of the numbers of sales.

This research was with the courtesy support of the local super market in Japan. We express appreciation to those involved.

References

1. Recommendation Engine White Paper, NetPerceptions (2000), http://www.netperceptions.com/literature/content/recommendation.pdf
2. Linden, G., Smith, B., York, J.: Amazon.com Recommendations; Item-to-Item Collaborative Filtering. IEEE Internet Computing, 73–80 (January-February 2003)
3. Orma, L.V.: Consumer Support Systems. Communications of the ACM 50(4), 49–54 (2006)
4. Hijikata, Y.: Techniques of Preference Extraction for Information Recommendation(Special Features-Exploiting Customer's Preference: Leading Edge of User Profiling Technique). Journal of Information Processing Society of Japan, Information Processing in Japan 48(9), 957–965 (2007)
5. Burke, R.: Hybrid Recommender Systems: Survey and Experiments. User Modeling and User-Adapted Interaction 12, 331–370 (2002)
6. Adomavicius, G., Tuzhilin, A.: Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. IEEE Trans. on Knowledge and Data Engineering 17(6), 734–749 (2005)
7. Herlocker, J., Konstan, J., Terveen, L., Riedl, J.: Evaluating Collaborative Filtering Recommender Systems. ACM Trans. on Information Systems 22(1), 5–53 (2004)
8. Schafer, J.B., Konstan, J.A., Riedl, J.: E-Commerce Recommendation Applications. Data Mining and Knowledge Discovery 5, 115–153 (2001)
9. Denning, P.J., Dunham, R.: The Missing Customer. Communications of the ACM 50(4), 19–23 (2006)
10. Nakamura, H.: Marketing of New Products, Chuokeizai-Sha (2001) (in Japanese)
11. Abe, M., Kondo, F.: Science of Marketing -POS data Analysis. Asakura Publishing (2005) (Japanese)
12. Takahashi, M., Nakao, T., Tsuda, K., Terano, T.: Generating dual-directed recommendation information from point-of-sales data of a supermarket. In: Lovrek, I., Howlett, R.J., Jain, L.C. (eds.) KES 2008, Part II. LNCS (LNAI), vol. 5178, pp. 1010–1017. Springer, Heidelberg (2008)
13. http://taste.sourceforge.net
14. Sarwar, B., Karypis, G., Konstan, J., Riedl, J.: Application of Dimensionality Reduction in Recommender System. In: ACM WebKDD Workshop (2000)

A Study on Comprehending the Intention of Administrative Documents in the Field of e-Government

Keiichiro Mitani[1], Yoshinori Fukue[2], and Kazuhiko Tsuda[1]

[1] Graduate school of Systems Management, University of Tsukuba, Tokyo
3-29-1,Otsuka,Bunkyo-ku, Tokyo 112-0012, Japan
mitani@gssm.otsuka.tsukuba.ac.jp
[2] FUJITSU LIMITED

Abstract. Policy plans published by government institutions have been widely available to the public on the Internet. However, it has been noted that there often exist many unclear expressions in the sentences. In this paper, we have proposed a method of considering the level of ambiguity as relates to the amount of confidence the writer has in the implementation of the policy presented. In the policy plans analyzed in this study, many unclear expressions existed at the end of sentences. We also confirmed that such policies using unclear expressions lowered the rate of successful implementation as a whole. We also analyzed the appearance frequency of the expression "nado (and so forth)" which makes nouns and verbs within sentences unclear. As a result, we were able to confirm that the more expressions such as "and so forth" are used within one sentence, the lower the success rate of implementation of the policy becomes.

Keywords: text mining; e-Government.

1 Background and Purposes

The rapid expansion of the Internet has provided vast amounts of literal computerized information. Finding significant knowledge from vast amounts of information is very important; however, it is almost impossible for humans to do that alone. In such an environment, the use of text mining is a highly effective method in order to automatically analyze document information.

However, automatic analysis is still inadequate when comprehending the writer's intention and as to his thoughts and feelings described within the document.

By using analogy in this study, we focused on the content of policy plans described by government institutions in order to examine methods to reason on the writer's intention for policy implementation. Although policy plans in Japan are widely available to the public through media such as the Internet, many individuals have indicated that unclear expressions are contained within the text. As to one of the causes of these ambiguous expressions, we considered that the level of writer's confidence could have an effect as to whether the policy stated in the document can be achieved or not. We examined whether the level of confidence could be viewed as an intention.

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 293–299, 2009.

Text mining is a method to understand the meaning of the relevant document through morphological analysis and syntax analysis. This study aims to gain a deeper comprehension of the writer's intention, by analyzing vague phrases and expressions.

2 Related Studies

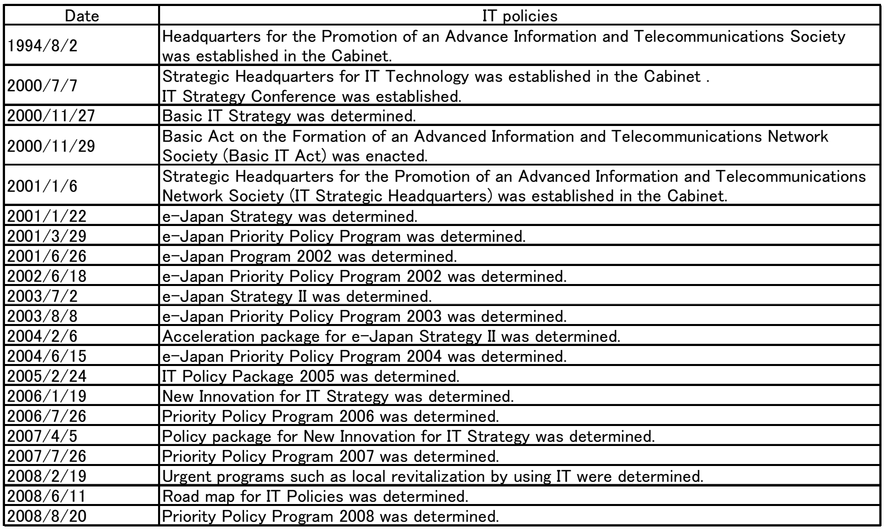
There has been some previous research work that aimed to comprehend the intention of writers by utilizing text mining. Takahashi has analyzed the relationship between targeted stock prices by reading intentions from text data described in analysis reports by analysts [1]. Suzuki has worked on estimating writer's intentions and feelings from emoticons that are frequently observed in colloquial descriptions made on the Internet [2].

In addition, some studies have been conducted to predict potential problems that may arise in the future by analyzing the content of documents. Murakami et al. have studied how predicting potential troubles that may appear in the future, by analyzing the content of project progress reports related to software development [3]. Takuma et al. have proposed a method that utilizes text data continuously stored in call centers in order to discover problems at an early stage [4].

3 About IT Policies in Japan

In Japan, the Basic Act for the Formation of an Advanced Information and Telecommunications Network Society was established in 2000. Given this act, IT Strategic

Table 1. Transition of Japan's IT policies



Date	IT policies
1994/8/2	Headquarters for the Promotion of an Advance Information and Telecommunications Society was established in the Cabinet.
2000/7/7	Strategic Headquarters for IT Technology was established in the Cabinet . IT Strategy Conference was established.
2000/11/27	Basic IT Strategy was determined.
2000/11/29	Basic Act on the Formation of an Advanced Information and Telecommunications Network Society (Basic IT Act) was enacted.
2001/1/6	Strategic Headquarters for the Promotion of an Advanced Information and Telecommunications Network Society (IT Strategic Headquarters) was established in the Cabinet.
2001/1/22	e-Japan Strategy was determined.
2001/3/29	e-Japan Priority Policy Program was determined.
2001/6/26	e-Japan Program 2002 was determined.
2002/6/18	e-Japan Priority Policy Program 2002 was determined.
2003/7/2	e-Japan Strategy II was determined.
2003/8/8	e-Japan Priority Policy Program 2003 was determined.
2004/2/6	Acceleration package for e-Japan Strategy II was determined.
2004/6/15	e-Japan Priority Policy Program 2004 was determined.
2005/2/24	IT Policy Package 2005 was determined.
2006/1/19	New Innovation for IT Strategy was determined.
2006/7/26	Priority Policy Program 2006 was determined.
2007/4/5	Policy package for New Innovation for IT Strategy was determined.
2007/7/26	Priority Policy Program 2007 was determined.
2008/2/19	Urgent programs such as local revitalization by using IT were determined.
2008/6/11	Road map for IT Policies was determined.
2008/8/20	Priority Policy Program 2008 was determined.

Headquarters in which the Prime Minister serves as the Director-General was established in 2001. In the same year, “e-Japan Strategy” that is Japan’s first comprehensive IT policy [5] was also established. In the present study, we conducted analyses based on the policy plan publicized as e-Japan Priority Policy Program [6] out of all the IT policies. Table 1 shows the transition of Japan’s IT policies.

The reason why we focused on this plan in this study is that the continuity of policy measures was easy to confirm, and success and failure of each policy measure was comparatively easy to judge. With respect to success and failure of individual policy measures, only those measures in public documents that have been considered to clearly be implemented with the goals achieved were regarded as implemented successfully.

4 Comprehension of Intension Regarding Sentence-Ending Expressions

To conduct our analysis we focused on the expressions used for the end of each sentence that described the policy measure. The following expressions are used as sentence-ending expressions: “… wo kento suru (to examine sth),” “… wo suishin suru (to promote sth),” or “… wo jisshi suru (to implement sth).” These Japanese expressions used at the end of sentences can show the writer’s attitude toward the policy measure, namely the level of writer’s confidence. If sentence-ending expressions are directly linked to the level of the writer’s confidence, then we should be able to predict that the sentence-ending expressions have a correlation with the rate of successful policy measures.

First, in the target document, we analyzed the type of sentence-ending expressions and their frequency of appearance. Fig.1 shows 21 types of sentence-ending expressions that have a 1% or more of appearance frequency observed in the entire document. On the whole, it seems that a significant bias exists in the sentence-ending

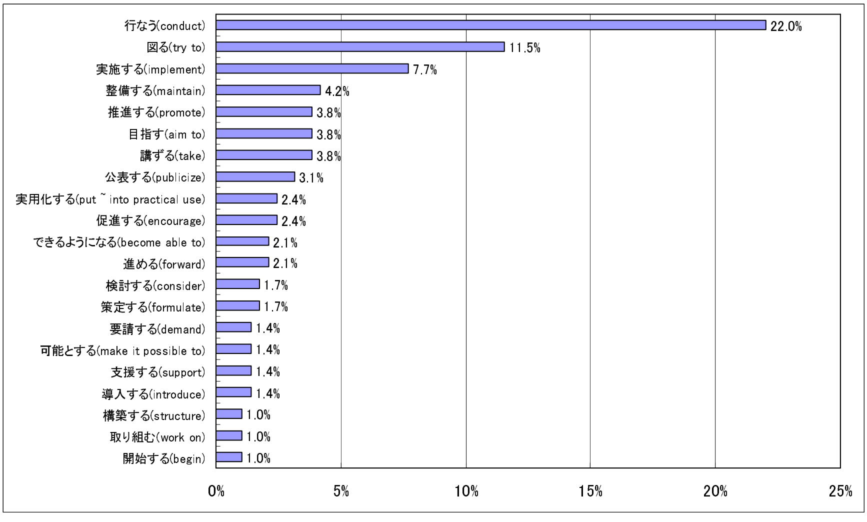


Fig. 1. Appearance frequency of each sentence-ending expression

expressions of administrative documents. The top 4 expression types cover a 40% of the total, and the top 10 expression types cover a 60% or more of the total.

The following sentence-ending expressions came to the top: “okonau (conduct sth)” “hakaru (try to)” “jisshi suru (implement sth)” and “seibi suru (maintain sth)”. As to the common characteristic of these expressions, they do not have a single meaning, but are verbs that can be widely used in various situations. The expression “okonau (conduct sth)” is a versatile expression which can connect to a wide variety of words. The expression “hakaru (try to)” has a meaning close to the word “examine” which means “to consider and judge,” as well as multiple meanings such as “to plan” that indicates “to attempt to” and “to contemplate,” and “to put into practical use” that indicates “to process” and “to arrange.” The expression “seibi suru (maintain sth)” was used in the target document as follows: to maintain law, to maintain the system, to maintain the environment, and to maintain the foundation. In government institutions, this expression seems to be used in wider concepts than used in general.

The uses of such sentence-ending expressions widely combined with multiple meanings are preferred. This might be because the government can easily avoid responsibility even if the implementation of a certain policy does not go as planned.

On the other hand, when the rate of successful policy measures of each sentence-ending expression (the ratio where the policy was implemented and the goal was achieved) was analyzed, the results are shown in Fig. 2.

It seemed that the rate of successful policy measures was dimidiate with regard to clear sentence-ending expressions that do not frequently appear. Those expressions such as “sakutei suru(formulate sth)” and “dounyu suru(introduce sth)” seemed to have a high success rate; however, other sentence-ending expressions with the meaning of having a high barrier for implementation, such as “kanou to suru(make it possible to)” “kouchiku suru(structure sth)” and “jitsuyouka suru(put sth into practical use)” seemed to have a low success rate.

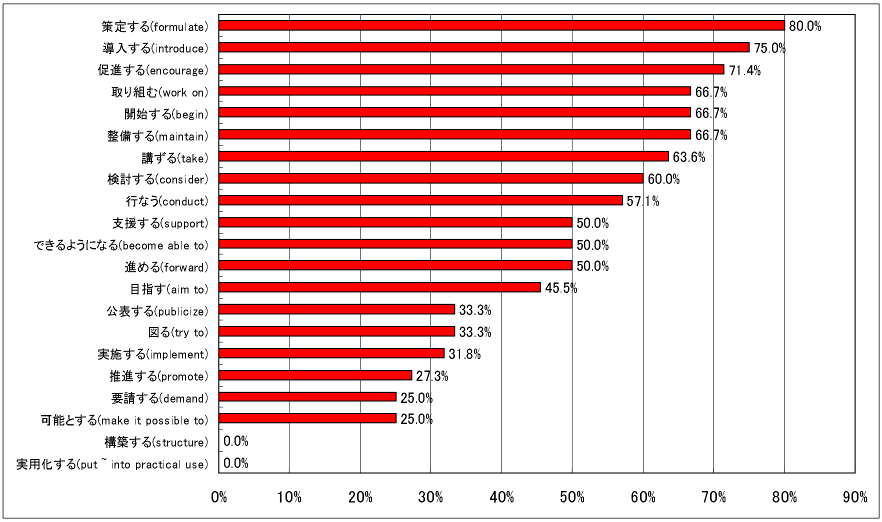


Fig. 2. Success rate of each sentence-ending expression

In contrast, regarding unclear sentence-ending expressions frequently used, the expression “okonau(conduct sth)” showed a success rate (57.1%) higher than the overall average (47%), whereas the expression “hakaru(try sth)” showed a 33.3% of success rate and the expression “jissi suru(implement sth)” showed 31.8%. These expressions showed very low success rates.

5 Comprehension of the Intention Regarding the Expression “and So Forth” (“等(nado)”in Japanese)

In administrative documents, the expression “and so forth” frequently appears compared to general documents. This expression is used as follows: We are going to visit the United States, Britain, France, and so forth. This expression is often used in order to indicate that there may be other similar things in addition to the identical things earlier listed. However, a frequent use of this expression may cause a risk that the meaning of sentence becomes more and more vague.

Table 2 shows the calculation and analysis results of the expression “and so forth” as it appeared in each chapter of the administrative document analyzed in the present study.

Applications of this expression can also be categorized in the following three ways.
[Noun application]: To obscure the meaning of nouns such as “the Internet and so forth,” “dispatched employees and so forth,” and “pharmacies and so forth.”
[Verb application]: To obscure the meaning of verbs such as “to introduce sth and so forth,” “to submit sth and so forth,” and “to review sth and so forth.”
[Other application]: Application other than the above ones such as “high-efficiency and so forth” and “fulfilling and so forth.”

The appearance rate of the expression “and so forth” in each chapter was between 0.7% and 1.2%, the average rate was 0.9%, and was applied as a noun 75% of the time.

The rate of successful policy measures including the expression “and so forth” was between 28% and 47%. The average was 36%. Since the average rate of all the successful policy measures was 47%, this fact shows that such policy measures applying

Table 2. Appearance rate of “and so forth”

	Advanced Information and Telecommunications Network	Cultivation of human resources	Promotion of e-commerce	Computerization of the administration	Total
Noun application	35	42	40	82	199
Verb application	11	13	14	19	57
Other application	1	2	3	2	8
Total	47	57	57	103	264
Total words	7,229	7,468	6,400	8,893	29,990
Appearance rate of “and so forth”	0.7%	0.8%	0.9%	1.2%	0.9%
Rate of successful implementation	47%	40%	37%	28%	36%

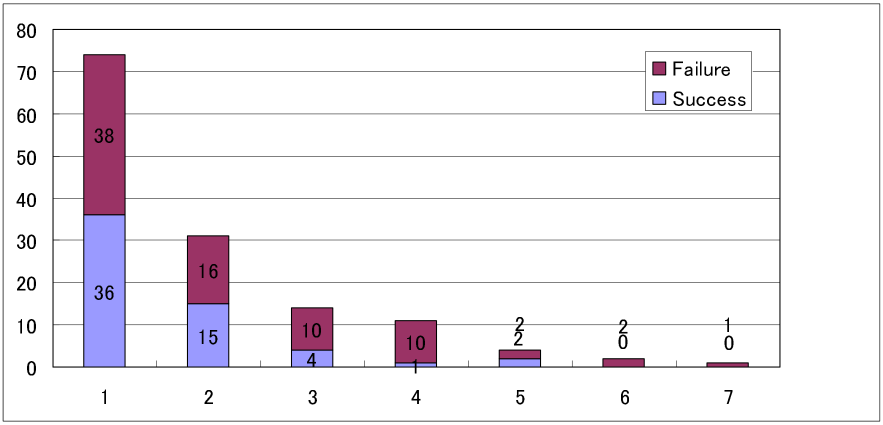


Fig. 3. No. of the expression "and so forth" and the success and failure of policy measure

the expression "and so forth" had the tendency to somewhat decrease the rate of successful implementation. In the present study also, the lowest success rate of 28% was recorded in the chapter, "Administrative Computerization," in which the expression "and so forth" most frequently appeared (1.2%).

In addition, the number of times this expression appeared within one sentence which indicated the policy measure was also focused on and analyzed. At the most, this expression was used seven times within one sentence.

When the rate of successful policy measures by the number of times the expression "and so forth" appeared, the rate was 48% when this expression appeared once, 48% for twice, 28% for three times, and 9% for four times. This result shows that the success rate tended to decrease as the number of appearance increases (see Fig. 3). Only in the case where this expression appeared five times was there showed a 50% of success rate; however, both of the policy measures succeeded in implementation were where the measure formulated a legal system, and the name of legal system itself included "and so forth." Thus such an exceptional result was produced. It seems that the more the sentence becomes obscure, the more the policy measure itself becomes unclear, and it ends up in failure without implementation.

6 Conclusion

In this study, we analyzed the content of policy plans described by the government in order to examine whether the writer's intension could be comprehended.

As a result of analyzing sentence-ending expressions contained in the document, we were able to confirm that there exists a correlation between sentence-ending expressions and the rate of successful policy measures. It would appear that a sentence including those unclear sentence-ending expressions frequently used tends to lower the rate of successful implementation. However, with regard to clear sentence-ending expressions, the success rate seems to broadly be divided in the one with a high success rate and the other one with a low success rate. Given this analysis, we can say that the level of the writer's confidence for the implementation of policy measures can be shown in sentence-ending expressions.

We were also able to confirm through analysis of the expression "and so forth" that there exists a high correlation between the number of times this expression is used and the implementation of policy measures. This expression is used in order to make the sentence meaning obscure. The more the document contains this expression, the more its concreteness becomes less. In other words, when a writer frequently uses this expression, it means the writer has no specific fixed image of the policy measure. This result shows that the level of writer's confidence can be reflected by using such unclear expressions.

It is appropriate to consider that the same kind of state can be observed in administrative institutions as well as ordinary businesses.

We are interested in examining numerical expressions that have meaning in order to increase sentence clarity in the future. We would like to continue this study aiming at comprehending the correlation between written expressions and the writer's intention with a higher degree of accuracy.

References

[1] Takahashi, S., Takahashi, H., Takahashi, M., Tsuda, K.: Analysis of the Validity of Textual Data in Stock Market through Text Mining. WSEAS Transaction on Business and Economics 3(4), 310–315 (2006)

[2] Suzuki, N., Tsuda, K.: Express Emoticons Choice Method for Smooth Communication of e-Business. In: Gabrys, B., Howlett, R.J., Jain, L.C. (eds.) KES 2006. LNCS (LNAI), vol. 4252, pp. 296–302. Springer, Heidelberg (2006)

[3] Murakami, A., Nakamura, T.: The prediction of the trouble which the surface expression of the project progress report was used for. In: The 22nd Annual Conference of the Japanese Society for Artificial Intelligence, pp. 1–4(3B1-2) (2008)

[4] Takuma, D., Nomiyama, H.: Early Problem Detection Using Text Data, Information Processing Society of Japan SIG Technical Report, pp. 19–26 (2004-NL-162) (2004)

[5] The advanced information communication network society promotion strategic headquarters (2001) e-Japan strategy, http://www.kantei.go.jp/jp/it/network/dai1/pdfs/s5_2.pdf

[6] The advanced information communication network society promotion strategic headquarters, The intensive plan for e-Japan strategy (2001), http://www.kantei.go.jp/jp/singi/it2/kettei/010329honbun.html

Decision Making Process for Selecting Outsourcing Company Based on Knowledge Database

Akihiro Hayashi, Yasunobu Kino, and Kazuhiko Tsuda

Graduate School of Business Sciences, University of Tsukuba,
3-29-1 Otsuka, Bunkyo-ku, Tokyo, 112-0012, Japan
{hayashi,tsuda}@gssm.otsuka.tsukuba.ac.jp,
kino@mbaib.gsbs.tsukuba.ac.jp

Abstract. In system development projects, there is increasing number of cases requiring decision making to be conducted using official process. In this paper, the idea of Analytic Hierarchy Process and Linear Programming are introduced into the framework of the traditional decision making process. In addition, decision making process is continuously improved by eliminating cognitive bias in evaluator, by registering the differences between the initial decision making result and corrected result to Knowledge Database. The proposal was applied to a real case and the width of correction required has reduced after series of process improvements.

Keywords: Decision making Process, Kepner-Tregoe Method, AHP Method, Linear Programming, Cognitive biases.

1 Introduction

Recently, there is increasing number of cases in system development projects where decision making (hereafter, abbreviated as DM) using official process is required. An example is selection of outsourcing companies (hereafter, abbreviated as OCs).

However, it is very difficult and unrealistic for the project managers to investigate advanced expertise of evaluation method like simulation model or probability model and establish process for selecting OCs within tight development schedule. As a result, there are few cases where evaluation method process implemented effectively. From these backgrounds, establishment of the decision making process for selecting OCs is taken up as the theme of this research.

As an early research on outsourcing of IT Management, Lacity [1] pointed out that the success condition of the outsourcing depends on the maximization of flexibility and the control ability of the system and discussed a concrete selection method for selecting OCs. Huber [2] listed various conditions of selecting OCs which change high fixed costs into the additional value ahead, then discussed a method for candidate organizations selection. Cross[3] reported on the means to evaluate the candidate company list of outsourcing under principle of competition in IT outsourcing strategy. However, there is no specific previous research that objectively evaluated the DM process for selecting OCs in system development projects.

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 300–307, 2009.

2 Issues to Be Solved in the Traditional DM Process

A traditional DM method is used to establish DM process. The traditional DM process has its root in Kepner-Tregoe Method [5] that is developed in the United States and is executed by following steps: 1)Define issue that is requiring decision,2)Set up the evaluation criteria for the alternatives, 3)Set relative weight of the evaluation criteria,4)Enumerate all alternatives, 5)Evaluate alternatives based on the criteria,6)Evaluate value combining criteria weight and each value,7)Select the highest score of the comprehensive evaluation.

The aspect of the traditional DM process is that in Step 2 where the criteria are weighted and in Step 5 where the alternatives are evaluated, numerical values are used in evaluation. Therefore, total score is semi automatically calculated.

2.1 Issue of Weighting Criteria

Weighting Criteria is that selection criteria are classified into absolute conditions and relative conditions, and weight is set to respective relative conditions. The absolute condition is a mandatory condition such that must be met. On the other hand, a relative condition is such that it is preferred to be met, but not an absolute condition.

The criteria in traditional DM method are usually evaluated by ten stages evaluation where score of 1-10 is set on each of the relative conditions. However, the rationale of the logical grounds is extremely vague. It might be difficult to reproduce the same point accurately, even when the same evaluator executes it twice.

Clear grounds to grade do not exist when one evaluates by 10 stages for the criteria in relative condition. This becomes the issue in weighting criteria.

2.2 Issue of Alternatives Evaluation

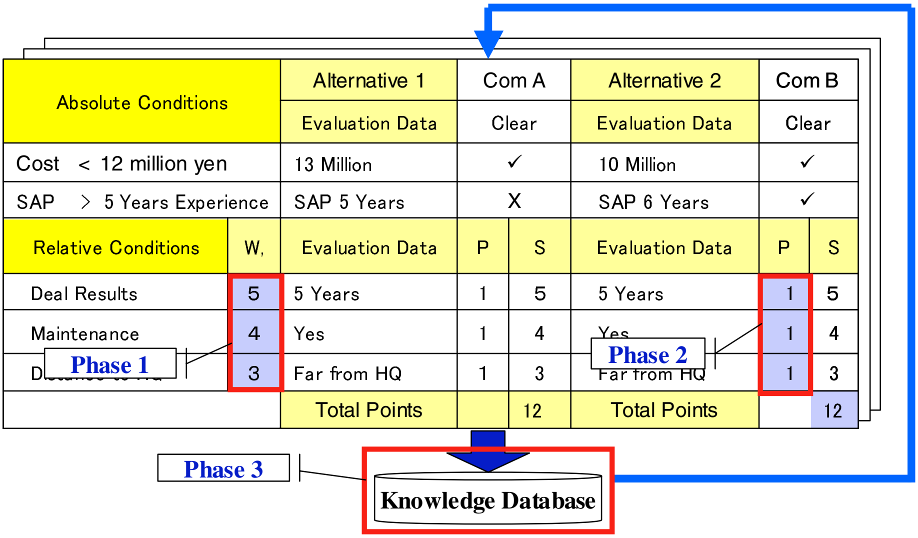
It is difficult for evaluator to objectively evaluate the alternatives numerically due to cognitive biases. Cognitive bias is the phenomenon that evaluation of certain object is dragged by a remarkable feature, or misinterpreted by influence from specific information.

For example, Halo Effect that the evaluation is unnecessarily improved due to introduction by the customer, Ranking Inflation that makes the evaluation lenient when the candidate has long term association, or Ranking Compression that the evaluation becomes noncommittal near center saying "It cannot be said either" [6].

The issue to be solved is that the alternatives are to be evaluated objectively without influence of cognitive biases.

3 DM Process for Selecting OC

A method to solve the issues of traditional DM process is proposed here. The method consists of three phases shown in Figure 1.



Absolute Conditions		Alternative 1	Com A		Alternative 2	Com B	
		Evaluation Data	Clear		Evaluation Data	Clear	
Cost < 12 million yen		13 Million	✓		10 Million	✓	
SAP > 5 Years Experience		SAP 5 Years	X		SAP 6 Years	✓	
Relative Conditions	W,	Evaluation Data	P	S	Evaluation Data	P	S
Deal Results	5	5 Years	1	5	5 Years	1	5
Maintenance	4	Yes	1	4	Yes	1	4
Distance to HQ	3	Far from HQ	1	3	Far from HQ	1	3
		Total Points		12	Total Points		12

Fig. 1. Decision Making Process for Selecting Outsourcing Company

Phase 1: Applying AHP for Weighting Criteria

In this paper, by using Analytic Hierarchy Process [7] (hereafter, abbreviated as AHP), relative conditions are weighted. AHP is a technique for choosing the best evaluation by synthesizing a relative importance of the element in each hierarchy after arranging them to a layered structure of target, evaluation criteria, and alternative approach.

Weight is set at an average value of a pairwise comparison. Point is set from the classification of (1) Same, (3) a little, (5) rather, (7) plentifully, (9) Absolutely (inverse is used when one calculates opposite way) [8].

The product of the evaluation result of the criteria is calculated and their geometric mean is determined. Then proportion of each geometric mean to the total of the geometric mean is considered to be the weight.

Phase 2: Evaluate Alternatives and Verification using Linear Programming

Alternatives are evaluated by criteria and an integrated point is calculated. Then the result is verified by Linear Programming (hereafter, abbreviated as LP). LP is a technique used to calculate the optimum.

The 1st and 2nd place alternatives of the result by 10 points evaluation are verified using LP. Evaluation point required for the 2nd place alternative to exceed point of 1st place alternative is calculated by LP method. Then, more than two people compare these values to verify that there was no effect of cognitive bias.

For instance, there is possibility to generate Cognitive biases due to the person in charge of the evaluation with insufficient understanding of either relative conditions, past incident of delivery delay and/or a quality trouble by OC in question.

In such cases, reevaluation is required having person in charge of the evaluation replaced to those who is unlikely to have cognitive bias. If there is difference between original and second evaluator, the value determined by "second evaluator's point minus original evaluator's point" is recorded as “Cognitive Bias Point”.

Phase 3: Optimization of Evaluation using Knowledge Database
Next, using the cognitive bias points obtained in 3.2, evaluation value is optimized. Cognitive bias point is such that when evaluator evaluates a relative condition against a specific alternative, evaluation value increases or decreases in constant direction by influence of constant cognitive bias. Examples of Cognitive Bias Points are such that;

- Evaluator-A evaluates the quality control level of Company-C 2 point lower.
- Evaluator-B evaluates the skill of SAP engineers in Company-D 3 point higher.

When the point obtained from LP is judged to be more appropriate compared to the initial point, then it can be said that there is cognitive bias in evaluation of relative condition of alternatives. Then, a matrix of evaluator, relative condition, and alternative is created, and the cognitive bias point is registered to Knowledge Database (hereafter, abbreviated as KDB).

When a similar case (same evaluator, relative conditions, and alternatives) occurred, the mean value of their cognitive bias points is registered, which is calculated by total cognitive bias points and execution frequency. At next occasion of DM, the evaluation value is adjusted appropriately by subtracting the cognitive bias points from the evaluated value to prevent cognitive bias.

4 Application Evaluation

In this chapter, effectiveness of the proposal in this paper to the DM process is evaluated by applying to a real system development project.

4.1 Case Study

The case used here is a scene of selecting an OC in system development project in IT vender ‘Company-N’ with CMMI Level 3.

The rationales for outsourcing to the external organization include; to focus on the core competence; to pursue cost advantage. However, despite outsourcing being promoted to achieve cost advantage, number of troubles such as delivery delay or quality issue, due to skill-shortage and lack of communication capability in OCs has occurred. The maximum budget of the project for outsourcing development is 12 million yen. System to be outsourced is the program development of SAP software operated on Windows Vista. A bridge engineer is to be assigned between Japan and offshore site, with strong expectation for fluent Japanese capability.

4.2 Application of the Proposal to Case

DM process of selecting OC proposed in this paper is applied to the case. Details are as follows.

1) Define the Policy of DM. Company-N is to pursue cost advantage. Cheaper outsourcing is crucial, but cost is not the only criteria for selecting OC as quality problem had occurred in the past when OC was selected solely by cost.

Hence, Company-N established the basic policy for OC selection, which is "Cost is primary criteria, but not the only criteria. OC with capability of Japanese language, knowledge of Japanese culture, and abundant experience in SAP is to be selected".

2) Establish Absolute and Relative Conditions as Criteria. Since project cannot have a deficit, upper bound for budget of project and outsourcing cost are fixed beforehand. Moreover, having abundant SAP experiences is indispensable to prevent similar quality issue. It is preferable that criteria are clear judgment criteria. Hence, numerical value was put into criteria and absolute conditions were set as follows.

Table 1. Establish Evaluation Criteria (Absolute Condition)

	Conditions
Abs. Criteria 1	Order cost is less the 12 million yen
Abs. Criteria 2	More than 5 years of SAP Experience

Also, as Company-N has experienced some delivery delay in offshore development in the past, number of years for maintenance contract, deal results, and capability to manage project using Japanese language only is defined in the policy of DM. To guarantee technical skill, number of SAP qualifications holders was set as relative condition.

Table 2. Establish Evaluation Criteria (Relative Condition)

	Conditions
Rtv Condtn 1	Years for Maintenance Contract
Rtv Condtn 2	Deal Results
Rtv Condtn 3	Capability to manage project using Japanese Language
Rtv Condtn 4	No. of SAP Qualification Holders

3) Weighting the Relative Criteria by AHP Method. Next, relative conditions shown in Table 2 are weighted and importance is set. To set importance to criteria, the AHP method is applied as described in 3.1.

	Main	Deal	Jap	SAP	Calculations	Product	GeoMean	Weight
Main	1	3	5	7	1*3*5*7	105.0	105.0	0.575
Deal	1/3	1	1/3	1/3	1/3*1*1/3*1/3	0.037	0.037	0.079
Japan	1/5	3	1	7	1/5*3*1*7	4.200	4.200	0.257
SAP	1/7	3	1/7	1	1/7*3*1/7*1	0.061	0.061	0.089
						Total	5.57	1.000

Fig. 2. Criteria Prioritization by AHP

4) Enumerate all alternatives. Company-N has already achieved CMMI Level3 (Staged Representation). CMMI is used for process management, and candidate of OCs is predefined as "Supplier's Candidate's List" or "List of Desirable Supplier" as defined in ‘Supplier Agreement Management’ process.

In the "Supplier's Candidate's List", entire candidate OCs are listed, which were selected by considering “Geographical location of the supplier”, “Supplier’s performance record on similar work”, “Engineering capabilities”, “Staff and facilities availability to perform work”, “Prior experience in similar applications”. Actually, there were only five candidates OCs that meet criteria, so all of these candidates became alternatives.

5) Evaluation of Alternatives against Absolute and Relative Conditions. Now, alternatives are evaluated based on selection conditions. There are five OCs alternatives in the case of Company-N. Firstly, absolute conditions are judged by ✓ X. Candidate OCs are requested to submit cost and SAP development years to Company-N. As a result, as two companies did not meet at least one of absolute requirements, hence evaluation result was X (NO GO) for both of them. Three remaining companies had met both requirements, evaluation result was ✓ (GO) respectively.

Alternatives that had met absolute conditions are evaluated against relative conditions. In evaluation of relative conditions, alternatives are compared and 10 point is given to the alternative that best meet relative condition. Next, 1-9 point is set to other alternatives by comparison with alternatives with 10 point. Therefore, 10 point is given to one alternative and similarly 1-9 points to other alternatives.

Having completed weighting relative conditions and evaluation of each alternative, score of each alternative was obtained by calculating product of evaluation point and weight of relative conditions. This was repeated for all relative conditions, and sum of these are considered to be the total score.

Figure 3 shows evaluation result in Company-N. At this point, alternative-A becomes the highest total score and hence it will be selected.

		Alt. A		Alt. B		Alt. C		Alt. D		Alt. E	
Abs	Cond	Eva	Cler	Eva	Cler	Eva	Cler	Eva	Cler	Eva	Cler
A 1	Cost	12	✓	12	✓	11	✓	13	X	11	✓
A 2	SAP	5	✓	6	✓	5	✓	6	✓	4	X
Rtv	weight	Eva	Sco	Eva	Sco	Eva	Sco	Eva	Sco	Eva	Sco
R 1	0.575	9	5.175	8	4.600	10	5.757				
R 2	0.079	7	0.553	10	0.790	6	0.474				
R 3	0.257	10	2.570	7	1.799	7	1.799				
R 4	0.089	7	0.623	10	0.890	6	0.534				
	Total	8.92		8.08		8.56					

Fig. 3. Evaluation for Relative Conditions and Total Points

6) Correction of Cognitive Bias using LP. Here, total score of the highest score alternative is 8.92, so evaluation result of a relative condition that yield total score of, say 9 point, is calculated by LP. Result is shown in Figure 4 (right most column).

Alternative C in figure 3 and the LP value in Figure 4 are compared. The value of alternative C and LP are same for relative conditions 1 and 3. Hence, for remaining alternatives 2 and 4, evaluator, contents of relative conditions and Company-C itself are investigated whether cognitive biases of “Negative Leniency” had any influence.

First of all, relative condition 4 ‘No. of SAP Qualification Holders’ is investigated. Number of qualification holders is evaluated here, and this is objective value and hence there is no room for cognitive bias to take place.

Next, relative condition 2 ‘Deal Results’ is investigated. There were three past engagements between Company-N and Company-C. No particularly significant “Negative Leniency” was found. On the other hand, Company-A with high evaluation result had a long term relation of more than 20 years, since company establishment. It seems that there are feelings that Company-A cannot be eliminated in selection. There is no “Negative Leniency” for Company-C, but “Halo effect” to Company-A. The evaluation of Company-C was relatively low, so the evaluation point was changed to 8 point as calculated by LP.

From this, total score of alternative C became 8.72 as a result. However, this value is still lower than of alterative A, and hence, alternative A is selected as final alternative.

(Net multiplication part of alternative C in Figure 4).

		Alternative A		Alternative C		Liner Program	
R.C	Weight	Evaluation	Score	Evaluation	Score	Evaluation	Score
1	0.575	9	5.175	10	5.757	10	5.750
2	0.079	7	0.553	8	0.474	8	0.659
3	0.257	10	2.570	7	1.799	7	1.766
4	0.089	7	0.623	6	0.534	9	0.824
	Total	8.92		8.72		9.00	

Fig. 4. Verification using LP

7) Corrective Action by use of KDB. Evaluation of alternatives is improved by registering result of the verification in KDB and reflecting it in future evaluation.

The conditions of corrective action of cognitive bias using LP are recorded, and matrix with evaluator, relative condition and alternatives, is created. With similar case (same evaluator, relative conditions, and alternatives) occurred, the mean value is registered, which is calculate by total cognitive bias points and execution frequency. For example, if cognitive bias point at next execution was +3 point, then total (5) is divided by executed frequency (2), and mean of 2.5 is registered in KDB.

For next time when Evaluator-A evaluates deal result of Company-C, this value is subtracted from evaluated result.

4.3 Evaluation

The proposal of in this study is evaluated by increase or decrease of correction frequency of the cognitive bias point in application to cases.

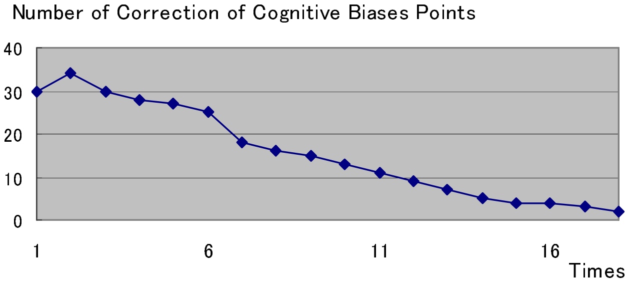


Fig. 5. Number of Correction of Cognitive Biases Points

When the cognitive bias is observed in DM process explained in 4.2, this is corrected. This correction should not be necessary from the next execution because the cognitive bias is subtracted and evaluate by use of KDB in similar cases.

Figure 5 is a correction transition of the cognitive bias points when this proposal is continuously applied for the DM process in Company-N. It can be observed that the value is corrected gradually, and correct decision is made without any noise.

5 Conclusion

In this paper, considering that traditional DM process is not sufficient to provide reliable DM process, application of AHP and LP is proposed. In the real case study in Company-N, the effect of cognitive bias in DM process was gradually decreased.

However, in order to declare that the DM process for selection of OC is successful, actual project should succeed in points of quality, cost, and delivery date, with contribution of OC to the success. Contribution to such a project result is not appreciable in this paper. This is a future task.

References

1. Lacity, M.C., Willcocks, L.P., Feeny, D.F.: IT Outsourcing: Maximize Flexibility and Control. Harvard Business Review Article (May 1995)
2. Huber, R.L.: How Continental Bank Outsourced Its Crown Jewels. Harvard Business Review Article (January 1993)
3. Cross, J.: IT Outsourcing: British Petroleum's Competitive Approach. Harvard Business Review Article (May 1995)
4. Kepner, C.H., Tregoe, B.B.: The New Rational Manager. Princeton Research Press (1981)
5. Globis Business Schoool, MBA Glossary, `http://gms.globis.co.jp/dic/`
6. Saaty, T.L.: The Analytic Hierarchy Process. McGraw-Hill, New York (1980)
7. Takeda, H.: Saaty no houhou niyoru Weight no Jak-kan no ginmi, AHP Jireishu, JUSE, pp.223–246 (1990)

Intelligent QA Systems Using Semantic Expressions

Yutaka Inada[1], Hideo Nakano[1], Shinkaku Kashiji[1], and Junichi Aoe[2]

[1] Institute of Laguage Understanding
1-32-1, Nakajousanjima, Tokushima-shi, Tokushima 770-0813, Japan
inada@inlaun.co.jp
[2] University of Tokushima

Abstract. In the man-machine interfaces, it is important to use dialogue understanding technologies. One of the practical application fields is a question and answering (QA) systems. In order to reply appropriate answers for user's questions, this paper presents a dialogue technique by transforming semantic expressions for both requests and answers. The measurements for the disrepute of the QA system are introduced for requests and answers, respectively. For the KAMOKUMA QA system generating answers which are reflecting user's intension, the presented scheme is applied. For the AQ data with 7,518 requests, the real time simulation to estimate user's sufficiency is computed.

Keywords: Question and Answering, Dialogue Systems, Semantic Expressions.

1 Introduction

In the man-machine interfaces, it is important to use dialogue understanding technologies. One of the practical fields is a question and answering (QA) systems, and it requires both the user's satisfaction related to the quality and the amount of questions to be managed. There are many QA researches for large databases, but they is no relation to CRM schemes [5] [6] [7]. The CRM researches include answering opinion questions by [15], good and bad expression understanding by [13], sentence subjectivity by [14], and estimating sentence types by [12].

One solution is to control the QA systems by human operators like a call center. In fact, operator can achieve response with good quality for users, but it is impossible to manage many questions because of high human expenses. A FAQ (Frequently Asked Questions) scheme is a well known approach that users can find good answers. It is difficult to determine whether the FAQ service is sufficient, or not. Hammond et al.[16] presented knowledge navigation of FAQ systems, but there was no discussion about evaluation of the FAQ service. The behavior of the FAQ scheme is similar to the QA system that takes questions as the input and replies answers because FAQ knowledge bases are static in general. Thus, this study is very important to analyze and to classify questions and answers of FAQ knowledge. Understanding approaches for the affective expressions [7] must be introduced, not text classification

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 308–316, 2009.

approaches by [8] [9] [10] [11]. Harada et al.[17] proposed another approach based on the levels of questions and answers for FAQ knowledge. In the approach, 1) Questions are classified by four types, IMPOSSIBLE, SIDE EFFECT, INSUFFICIENT and UNCLEAR, and the degree for each type is defined. 2) Measurement of kindness for solutions is defined by classified answers. Answers are classified by four types, ACTION, CONFIRMATION, EXPLANATION, and NO PROBLEM, and the degree is defined. 3) Measurements of sufficiency for the whole FAQ service are introduced by the 1) and 2).

This paper presents an intension understanding scheme and the measurement of disrepute of the QA system by extending the method of Harada et al.[17]. Questions are classified by four types based on the strength of request. Answers are classified by five types based on the politeness of answers. By using the presented method, the QA system can generate answers reflecting user's intension. This approach is evaluated by the AQ data with 7,518 natural language requests. The proposed scheme is applied to the KAMOKUMA QA that can generate answers reflecting user's intension.

2 QA Systems Based on Semantic

2.1 The Construction of QA Systems

The concepts of the QA system to be discussed here is to take semantic expressions for the input and to generate the corresponding semantic expressions as shown in **Figure.1**. The input semantic expressions must reflect on user's requests, or problems, and the output must include information to resolve the problems. Although the scheme of Harada et al.[17] is restricted to the FAQ systems, the aim of the presented method is to reflect politeness of more general QA systems.

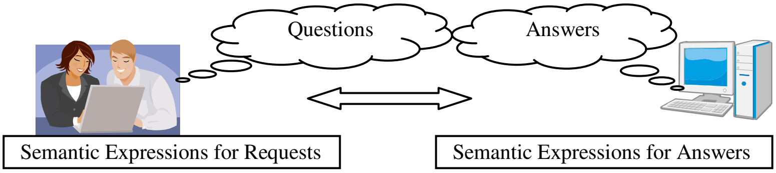

Questions
Answers
Semantic Expressions for Requests
Semantic Expressions for Answers

Fig. 1. An Illustration of QA Systems

2.2 Semantic Expressions [17]

A questioner is classified into three kinds of types: interrogative, imperative and declarative sentences. Consider each sentence requesting a drink as follows:

1) Interrogative sentence: "Don't you get a drink?"
2) Imperative sentence: "Give me a drink."
3) Declarative sentence: "I want a drink."

Although the above examples have direct intentions requesting a drink, many questions have indirect intentions. This is called an indirect speech act. Consider a question (1) "Aren't you thirsty?" with indirect intention. "How about some juice?" is one of the right answers if a questioner's intention is "I want a drink." The intention understanding depends on the dialogue situation and the semantic expression with a situation attribute must be formalized.

The stated indirect intention of question (1) = “Aren’t you thirsty?” can be represented by [Moisture is insufficient in the situation C], and the more formal description of a question semantic expression is denoted by [[[C],[SITUATION]];[[moisture],[OBJECT]]; [[insufficient], [CLAIM]]], where [] specifies semantic representations and [A] of [[A],[B]] is the attribute value for attribute [B]. In order to define an answer semantic expression corresponding to expected answer (a) for question (q), the question semantic expression is transformed by replacing [[insufficient],[CLAIM]] into [[supply], [SOLUTION]]. The transformed semantic expression is the answer semantic expression and it is represented as follows:

[[[C],[SITUATION]];[[moisture],[OBJECT]]; [[supply], [SOLUTION]]].

However, suppose that [C] is [the questioner with a juice and the respondent with no juice]. In this situation, the following dialogue is very strange.

Question (1) = “Aren’t you thirsty?” , Answer (a) = “How about some juice?”

One of right answers is “Please”. Therefore, the following semantic expressions are defined. A question semantic expression = [[[the questioner with a juice and the respondent with no juice],[SITUATION]]; [[moisture],[OBJECT]];[[sufficient],[REQUEST]]]. An answer semantic expression = [[[the questioner with a juice and the respondent with no juice],[SITUATION]];[[moisture],[OBJECT]];[[supply],[SOLUTION]]]. That means that, the questioner knows that the respondent has no juice and that he/she hopes to give his/her juice to the respondent.

2.3 Transforming Semantic Expressions

In the dialogue, a questioner provides useful answers resolving his/her requests, but it is difficult to reply good answers for any questions. Therefore, this section discusses a formal definition for dialogue systems by defining a Q-CLASS attribute [17]. The Q-CLASS attribute means the degree of questioner’s requests and it is defined by four kinds of levels.

We can define the degree of user’s requests by using Q-CLASS. By introducing additional attributes, the formal description of semantic expressions for questions and answers are as follows:

[Definition 1]. The question semantic expression SEMANTIC(*p*) for question *p* is defined as [C(SITUATION); x(OBJECT); i(CLAIM); a(Q-CLASS)].

Consider question *p* =“The printed character is unclear”. The question semantic expression is [[[printing], [SITUATION]]; [[character], [OBJECT]]; [[unclear], [CLAIM]]; [[INSUFFICIENT], [Q-CLASS]]].

[Definition 2]. The answer semantic expression for question semantic expression SEMANTIC(*q*) is defined as TRANSFORM(SEMANTIC(*p*)).

Consider question semantic expression SEMANTIC(*p*) = [[[printing], [SITUATION]]; [[character], [OBJECT]]; [[unclear],[CLAIM]];[[INSUFFICIENT],[Q-CLASS]]]. It means that [the user hopes that the printed character becomes clear] and one of the expected answers should be [recommend change the cartridge]. The answer semantic expression TRANSFORM(SEMANTIC(*p*)) becomes [[[printing], [SITUATION]]; [[cartridge], [OBJECT]]; [[change], [SOLUTION]]; [[ACTION], [A-CLASS)]]]. We can define the degree of answers by using A-CLASS.

[Definition 3]. For the semantic expression *r*, SURFACE(*r*) defines a set of surface sentences.

Consider the question semantic expression *r* = [[[printing], [SITUATION]]; [[cartridge], [OBJECT]]; [[change], [SOLUTION]]; [[ACTION], [A-CLASS)]]], SURFACE(*r*) includes sentence “The printed character is unclear”. By the same manner, for the answer semantic expression *r* = [[[printing], [SITUATION]]; [cartridge], [OBJECT]]; [[change], [SOLUTION]]; [[ACTION], [A-CLASS]]], SURFACE(*r*) includes “Please change a cartridge”**.**

Figure 2 shows an illustration of understanding process for the question “The printed characters are unclear” and the answer semantic expression.

In this case, value [printing] of attribute SITUATION is kept as the same value in the answer semantic expressions, but there are many semantic expressions to be changed. Consider question *p’*= “Paper is got blocked in printing”

SEMANTIC(*p’*) = [[[printing], [SITUATION]]; [[paper], [OBJECT]]; [[block], [CLAIM]]; [[SIDE EFFECT], [Q-CLASS]]] and TRANSFORM(SEMANTIC(*p’*)) includes [[[paper setting], [SITUATION]]; [[sheet holder], [OBJECT]]; [[fixed],

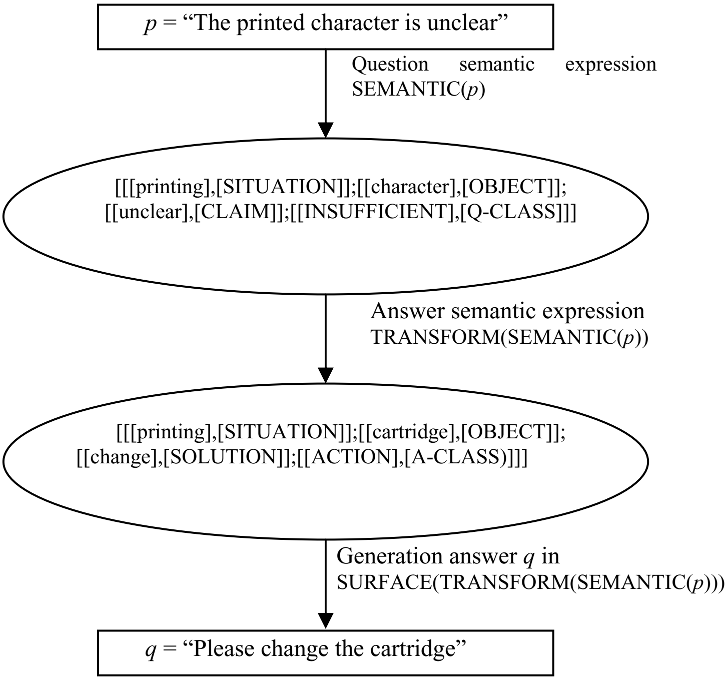

p = “The printed character is unclear”
Question semantic expression SEMANTIC(*p*)
[[[printing],[SITUATION]];[[character],[OBJECT]];
[[unclear],[CLAIM]];[[INSUFFICIENT],[Q-CLASS]]]
Answer semantic expression TRANSFORM(SEMANTIC(*p*))
[[[printing],[SITUATION]];[[cartridge],[OBJECT]];
[[change],[SOLUTION]];[[ACTION],[A-CLASS)]]]
Generation answer *q* in SURFACE(TRANSFORM(SEMANTIC(*p*)))
q = “Please change the cartridge”

Fig. 2. Transformation of semantic expressions

[SOLUTION]]; [[ACTION], [A-CLASS]]] as one of answer semantic expressions. This case means the situation of question *p'* is focusing on the detailed situation in the answer.

3 Degree of Disrepute for QA Systems

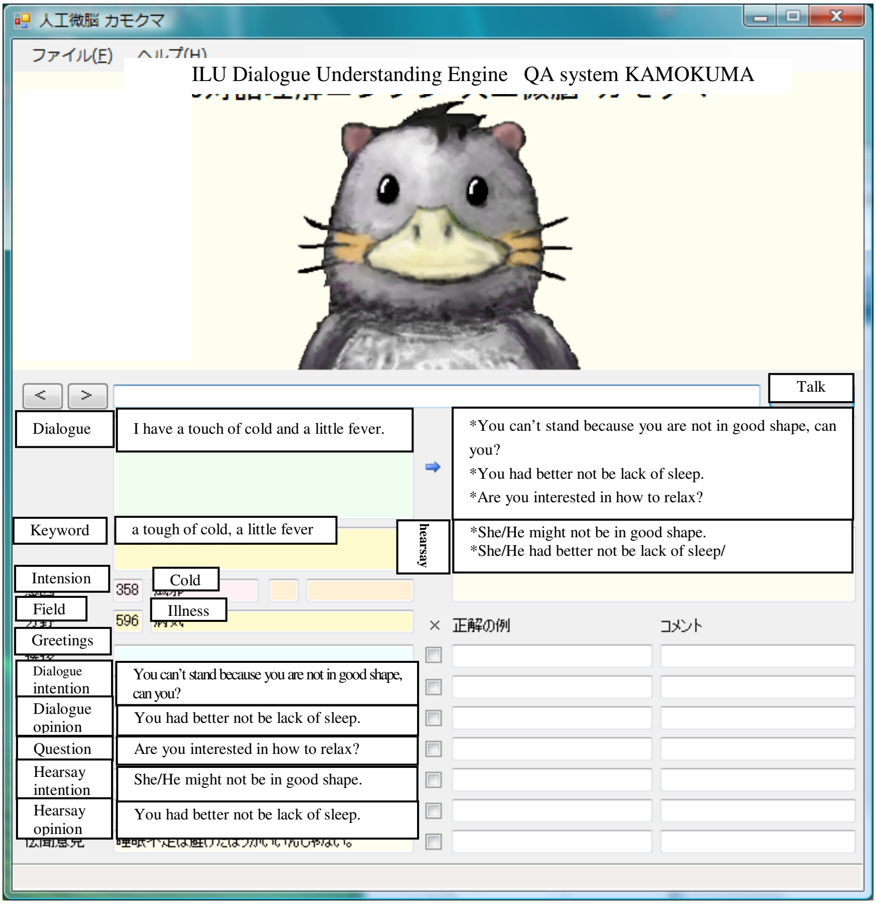
The emotion intension for requests is defined as 81 classes as shown in Figure 3. By using classes of **Figure 3**, the following Q-CLASS is defined.

1)glad	Reputation	2)good news	Reputation	3)lucky	Reputation
4)happy	Reputation	5)relieved	Reputation	6)celebrate	Reputation
7)grateful	Reputation	8)impressed	Reputation	9)satisfied	Reputation
10)comfortable	Reputation	11)delicious	Reputation	12)satisfied with the effect	Reputation
13)satisifed with the price	Reputation	14)pleasure	Reputation	15)expectation	Reputation
16)enjoyable	Reputation	17)pretty	Reputation	18)laugh	Reputation
19)entertainment	Reputation	20)compliment	Reputation	21)like	Reputation
22)encourage	Reputation	23)in good shape	Reputation	24)popular	Reputation
25)quick support	Reputation	26)kind support	Reputation	27)compliment for the support	Reputation
28)clear explanation	Reputation	29)good	Reputation	30)anger	Complaint
31)dressing-down	Complaint	32)dissatisfaction	Complaint	33)dishonor	Complaint
34)animosity	Complaint	35)criticism	Complaint	36)uncomfortable	Complaint
37)slander	Complaint	38)brackish	Complaint	39)dissatisfied with the effect	Complaint
40)poor support	Complaint	41)unkind support	Complaint	42)dissatisfied with the support	Complaint
43)dissatisfied with the price	Complaint	44)unclear explanation	Complaint	45)no response	Complaint
46)angriness	Complaint	47)sad	Complaint	48)bad news	Complaint
49)unlucky	Complaint	50)resignation	Complaint	51)lamentable	Complaint
52)disappointment	Complaint	53)apology	Complaint	54)lonely	Complaint
55)pity	Complaint	56)regret	Complaint	57)shocked	Complaint
58)worry	Complaint	59)rough	Complaint	60)sadness	Complaint
61)fear	Complaint	62)anxiety	Complaint	63)dislike	Complaint
64)bothered	Complaint	65)in bad shape	Complaint	66)unpopular	Complaint
67)abuse	Complaint	68)bad	Complaint	69)surprised	Others
70)request	Request	71)demand	Request	72)suggestion	Request
73)query	Question	74)inquiry	Question	75)application for membership	Reputation
76)application for withdrawal	Complaint	77)want to buy	Reputation	78)don't want to buy	Complaint
79)sold	Reputation	80)not sold yet	Complaint	81)invitation	Others

Fig. 3. Classes of request intensions

Table 1. Measurement DISREPUTE(Q-CASS, A-CLASS)

Q-CLASS A-CLASS	LEVEL 4	LEVEL 3	LEVEL 2	LEVEL 1
LEVEL 4	0(No dissatisfaction)	1(Overreaction, A little dissatisfied)	2(Missing points, Very dissatisfied)	4(Failure of Intension understanding, Terribly dissatisfied)
LEVEL 3	1(No concrete instruction, Very dissatisfied)	0(No dissatisfaction)	1(Missing points, A little dissatisfied)	3(Failure of Intension understanding, Strongly dissatisfied)
LEVEL 2	2(Only feedback, A little dissatisfied)	0(No dissatisfaction)	1(Missing points, A little dissatisfied)	2(Failure of Intension understanding, Very dissatisfied)
LEVEL 1	3(Strongly dissatisfied)	1(A little dissatisfied)	1(Missing points, A little dissatisfied)	2(Failure of Intension understanding, Very dissatisfied)
LEVEL 0	4(Terribly dissatisfied)	3(Strongly dissatisfied)	2(Strongly dissatisfied)	0(Satisfied)



Dialogue: I have a touch of cold and a little fever.

*You can't stand because you are not in good shape, can you?
*You had better not be lack of sleep.
*Are you interested in how to relax?

Keyword: a tough of cold, a little fever

*She/He might not be in good shape.
*She/He had better not be lack of sleep/

Intension: 358 Cold

Field: 596 Illness

Greetings

Dialogue intention: You can't stand because you are not in good shape, can you?

Dialogue opinion: You had better not be lack of sleep.

Question: Are you interested in how to relax?

Hearsay intention: She/He might not be in good shape.

Hearsay opinion: You had better not be lack of sleep.

Fig. 4. The KAMOKUMA QA system

(1) Q-CLASS = [LEVEL 4] :

For example, “I have a high fever because of the cold. (Complaint)” means a very serious claim, then it is defined as Q-CLASS = [LEVEL 4].

(2) Q-CLASS = [LEVEL 3] :

For example, “I think I have a cold. (Request) ” means a general claim (request), then it is defined as Q-CLASS = [LEVEL B].

(3) Q-CLASS = [LEVEL 2] :

For example, “Can I stop my cough? (Question) ” means weak claims(questions), then it is defined as Q-CLASS = [LEVEL 2].

Q-CLASS = [LEVEL 1] :

For example, “My cold got better. Thank you. (Satisfaction) ” means satisfaction, then it is defined as Q-CLASS = [LEVEL 1].

Suppose that the input is “I have a high fever because of the cold. (Complaint)”. Then, an A-CLASS attribute is defined as 4 kinds of levels.

A-CLASS = [LEVEL 4]

For example, “There is ABC hospital near here. The phone number is 123-456-789.” suggests a concrete solution, then it is defined as A-MODE = [LEVEL 4].

A-CLASS =[LEVEL 3]

For example, “Please take your temperature. Do you cough?” means actions, then it is defined as A-CLASS = [LEVEL 3].

(3) A-CLASS = [LEVEL 2]

For example, “My temperature is almost over 40 degrees C. I had better go to the hospital.” means a just explanation, then it is defined as A-CLASS = [LEVEL 2].

(4) A-CLASS = [LEVEL 1]: For example, “What should we do?” means no concrete solution, then it is defined as A-MODE = [LEVEL 1].

(5) A-CLASS = [LEVEL 0] : For example, “That’s good.” means agreement for satisfaction, then it is defined as A-MODE = [LEVEL 0].

By using Q-CLASS and A-CLASS, the measurement of disrepute for QA systems can be defined by DISREPUTE(Q-CASS, ACLASS) as shown in Table 1, where definition depends on the above examples.

4 Experimental Observations

In the simulation, the natural language analyzer with retrieving a variety of dictionaries has been utilized by using many techniques [1] [2] [3] [4]. The KAMOKUMA QA system by the presented method has been developed. **Figure 4** shows the demonstration picture.

For the KAMOKUMA system, 7,518 questions have been prepared for six kinds of requests for foods(1,367questions), trips(1,243 questions), health(1,082 questions), beauty(1,355 questions), products(1,234 questions) and sports(1,237 questions). **Figure 5** shows the results of DISREPUTE(R-CASS, ACLASS).

From the simulation results, the DISREPUTE can be utilized to take directly the tendency of user’s disrepute for QA systems. Intension understanding is very difficult techniques, but it turns out that the KAMOKUMA system has the understanding rate from 52% to 72%. The presented method can be applied to surveillance of the whole QA service using natural language processing schemes.

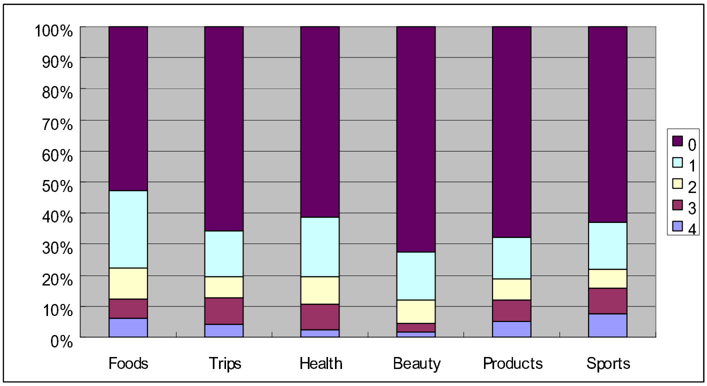


Fig. 5. The results of DISREPUTE(R-CASS, ACLASS)

5 Conclusions

This paper has been presented an estimation method of the QA service by introducing the following measurements: 1) user's disrepute for requests which is defined by four types of classifying questions and by four types of classifying answers. The formal measurement of disrepute in the QA systems has been defined.

The presented approaches have been evaluated by the KAMOKUMA QA system with 7,518 questions. Moreover, the real time simulation to estimate user's sufficiency (disrepute) has been computed. From this evaluation, it turned out that the presented approach is useful and effectiveness.

Acknowledgements. *The part of this research is research is supported by NICT*(National Institute of Information and Communications Technology)

References

[1] Aoe, J.: An efficient digital search algorithm by using a double-array structure. IEEE Trans. Softw. Engr. SE-15(9), 1066–1077 (1989)

[2] Fuketa, M., et al.: A document classification method by using field association words. An Inter. J. of Inf. Sci. 126(1), 57–70 (2000)

[3] Atlam, E.-S., et al.: Automatic Building of New Field Association Word Candidates Using Search Engine. Inf. Proc. & Manag (IPM) 42(4), 951–962 (2006)

[4] Atlam, E.-S., et al.: Documents similarity measurement using field association terms. IPM 39, 809–824 (2003)

[5] Kadoya, et al.: A Sentence Classification Technique by Using Intention Association Expressions. Computer Mathematics 82(7), 777–792 (2005)

[6] Ferret, L., et al.: QALC: the Question-Answering system of LIMSI-CNRS. In: The Ninth Text Retrieval Conference (TREC-9), pp. 235–244 (2001)

[7] Fukumoto, et al.: Question Answering Challenge (QAC1) Question answering evaluation at NTCIR Workshop 3. In: Third NTCIR Workshop Meeting: QAC1, pp. 1–10 (2002)

[8] Pang, B., et al.: Thumbs up? Sentiment Classification using Machine Learning Techniques. In: EMNLP, pp. 79–86 (2002)

[9] Kwon, O., Lee, J.: Text categorization based on k-nearest neighbor approach for Web site classification. In: IPM, vol. 39(1), pp. 25–44 (2003)

[10] Lam, W., Ruiz, M., Srinivasan, P.: Automatic Text Categorization and Its Application to Text Retrieval. IEEE Trans. on KDE 11(6), 865–879 (1999)

[11] Moens, M., Uyttendaele, C.: Automatic Text Structuring and Categorization as a First Step in Summarizing Legal Cases. IPM 33(6), 727–737 (1997)

[12] Tokunaga, H., et al.: Estimating sentence types in computer related new product bulletins using a decision tree. Information Sciences 168(1-4), 185–200 (2004)

[13] Fuketa, M., et al.: A Method of Extracting and Evaluating Good and Bad Reputations for Natural Language Expressions. Infor. Tech. & Deci. Making 4(2), 177–196 (2005)

[14] Hatzivassiloglou, et al.: Effects of adjective orientation and readability on sentence subjectivity. In: COLING, pp. 299–305 (2000)

[15] Yu, H., et al.: Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Sentences. In: EMNLP, pp. 129–136 (2003)

[16] Hammond, et al.: FAQ Finder: A Case-Based Approach to Knowledge Navigation. In: CAIA, pp. 80–86 (1995)

[17] Harada, J., et al.: Estimation of FAQ Knowledge Bases by Using Semantic Expressions for Questions and Answerd. International Journal of Computer Application in Technology 32(1), 69–81 (2008)

The Effective Extraction Method for the Gap of the Mutual Understanding Based on the Egocentrism in Business Communications

Nobuo Suzuki[1] and Kazuhiko Tsuda[2]

[1] KDDI Corporation
Iidabashi 3-10-10, Chiyoda, Tokyo 102-8460, Japan
nu-suzuki@kddi.com
[2] Graduate School of Buisiness Sciences, University of Tsukuba
Otsuka 3-29-1, Bunkyo, Tokyo 112-0012, Japan
tsuda@gssm.otuka.tsukuba.ac.jp

Abstract. The prospect of widespreading the Internet in business world is enabling new ways of solving our several business problems with many Q&A web sites. It is hard to say that good communication is carried out in these sites, because a lot of speakers may speak inconsistently in same site at same time. Therefore, we extracted the egocentrism from language expressions in words at some Japanese Q&A sites and tried to extract the gap of the mutual understanding in these words of business based on the egocentrism in this study. Specifically, we defined the weights for properties of presumed egocentrism with the method that presumes the egocentrism from language expressions in our previous study. We can presume the gap of the mutual understanding in the text data of business Q&A sites that have dialogue form by using the weights and calculating the strength score of the egocentrism in speech unit. We also evaluated this method with text data of business world in real Q&A sites and confirmed its effectiveness.

Keywords: Egocentrism, Gap of the mutual understanding, Q&A site.

1 Introduction

In recent years, the prospect of widespreading the Internet in our business society is enabling new ways of solving our several business problems with many Q&A web sites. Many companies utilize such sites to support their products to their users and improve the operational efficiency and the customer services. For example, OKWave is one of the famous Japanese Q&A sites for company's user support functions. Since such sites have a role as a field to resolve problems and an accumulating method to gather the collective intelligence, it is expected that extraction knowledge from them by the analysis using computational processing. However, it is hard to say that a good communication is performed in these sites, because a lot of speakers may speak individually in the same site at the same time [1]. This problem also causes difficulties of the automatic knowledge extraction. And managing sites itself may be exposed to a crisis. Therefore, in this study, we extracted the egocentrism from

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 317–324, 2009.

language expressions in words at some Q&A sites and tried to extract the gap of the mutual understanding in these words of businesses based on the egocentrism.

Here, the egocentrism means a mental disease that a person cannot throw away his/her own viewpoint and is the concept proposed by a child psychologist, Piaget [2]. The egocentrism also is a property of a human being that appears strongly in the days of an infant and weakens its skill in the communication with others with their growth. Futhermore, it is an obstruction factor of good communications on the Internet, and we have experienced it in daily life such as one-sided speech in e-mails [3]. For example, we may get an advertisement not a solution even if we request someone to teach something. Therefore, tools and methods are strongly desired to reduce such gap of the mutual understanding of their intentions.

In this study, we find properties of the egocentrism from Japanese sentences and show that it is available to express the gap of the mutual understanding by superficial verbalization without handling the deep meaning of sentences. Specifically, we define weights for properties of presumed egocentrism by the method of presuming the egocentrism with language expressions which was proposed in our former study. This method presumes the gap of the mutual understanding for text data in Q&A sites that have a dialogue form by calculating strength scores for the egocentrism in a speech unit with these weights. We also evaluated by the experiment with the text data in real Q&A sites and confirmed an effectiveness of our method.

2 The Egocentrism and Q&A Sites

There are some studies of Weblogs that are focused to the egocentrism so far. First of all, Numa et. al. defined the closeness from oneself with the distance based on degrees of the similarity of words and FOAF(Friend of a Friend) as the egocentrism, and proposed an information retrieval technique with it [4]. Next, Matsuoka et. al. modeled real human relations based on the Web link information such as meta data and track back data that centered on itself from Weblogs [5]. In contrast, these activities, our study places the egocentrism with real human emotional property not relations between Web sites, and extracts the gap of the mutual understanding in the dialogue with its language expressions.

On the other hand, we can find some researches that analyze features of dialogues from text data in Q&A sites as follows. Shimada et. al. related the fragmented articles by the relation of contributors [6]. Moreover, Murata et. al. extracted each community of the user and the question with the network that had the top of meta data of articles [7]. Aramaki et. al. extracted relations between comments with responding expressions and qualitative relation words [8]. These researches extracted the structural relationship between articles with only meta data not contents. In contrast, we extracted the semantic relationship of the gap of the mutual understanding in the dialogues by analyzing contents of articles in this study.

3 The Expression of the Egocentrism

The egocentrism is a basic and an emotional property of a human and appears on the text in many e-mails and Web sites. We collected many such language expressions,

Table 1. The classifications of the egocentrism and their weights

Classification	Subdivision	Weight
Attack to Others	Dissatisfaction	12
	Contempt	11
Lack of Empathy	Detachment	10
	Irony	9
Priority to One's Convenience	Defiant Attitude	8
	Restriction	7
Self Validity	Conclusion	6
	Imposition	5
	Excuse	4
Belonging to Self-Profit	Demand	3
Inference	Inference	2
	Question	1
Agreement	Thanks	0

and proposed the egocentrism classification and presumption method with n-gram morphemes analysis in our previous study [9]. In this study, we presume the gap of the mutual understanding in Q&A sites with this technique. Therefore, we have defined weights that show the strength for the classification of the egocentrism in table 1. The egocentrism is so strong and the gap of the mutual understanding is so large that the value of weight is large.

Here, we considered following points in addition to the previous research to apply this study. First, we added the classification of "Agreement" that was not included in the previous research. It has a particularly important meaning in dialogues and expresses how good mutual understanding goes. We extracted and classified the language expressions classified into "Questions" with only the question mark so far. But those expressions may not only show simple doubt, but also some opinions in the mutual understanding. Therefore, we excluded the nouns from objects of this analysis as simple questions if the nouns in the sentence of the question were repeated in the sentence of the answer.

Then, we calculated the weight of the egocentrism in a speech unit with the following method. First, we acquire the text data from Q&A sites and carry out the morphological analysis. If one speech consists of one or more sentences and the last sentence in the speech is the expression of the egocentrism, the egocentrism strength of the last sentence in a speech is defined as l. This is because the last sentence in a speech almost expresses the whole speech in many cases. Moreover, we decided the speech classifying to "Agreement" if it had at least one sentence expressed such meaning of an agreement. This is because sentences followed by the agreement are almost the additional explanation if it has at least one sentence of the agreement. On the other hand, if the last sentence doesn't have the expressions of the egocentrism, we calculate the mean of the strength x_i of each egocentrism for N sentences include the egocentrism in the speech. At last, as shown in Eq. (1), we get score S to represent the egocentrism of the speech with a larger value of this *l* and the mean.

$$S = \max(l, \frac{\sum x_i}{N}) \tag{1}$$

4 The Expression of the Gap of the Mutual Understanding in Dialogues

In this chapter, we explain the method to express the gap of the mutual understanding at the dialogues in Q&A sites as examples of business communication with the strength of the egocentrism in each speech calculated at last chapter. First, we plot the strength of the egocentrism of each speech at last chapter with a series of articles in each question and answer. We use to call these series of articles threads. Next, it is possible to presume the gap of the speech by the classification of the shape of the graph with classifying four patterns below. We used ChaSen as the morpheme analysis tool for Japanese to get morphemes [10]. It is the most common morpheme analysis tool in Japan.

(Pattern 1) The strength of the last speech is zero

The questioner can get an appropriate answer for the question and satisfaction finally in this pattern. This shows the gap of the mutual understanding in the dialogue is small (Fig. 1).

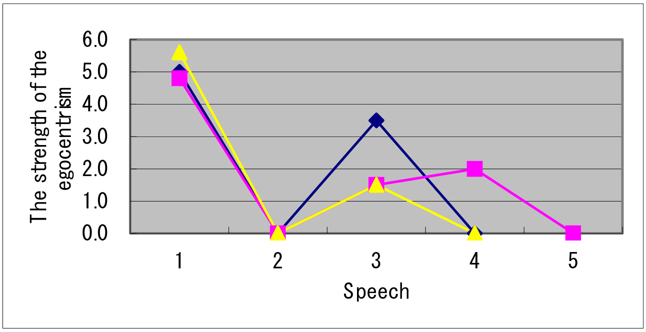


Fig. 1. The example that the strength of the last dialogue is zero

(Pattern 2) The strength of the speech just before the last is zero

In this pattern, the respondent can get the agreement for the answer of the question. Then, the additional explanation is provided. This also shows the gap of the mutual understanding is small (Fig. 2).

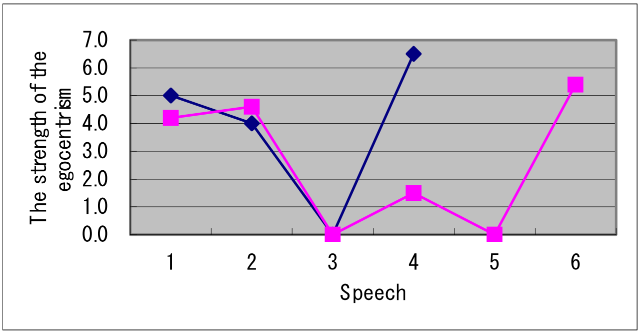


Fig. 2. The example that the strength of the speech just before the last is zero

(Pattern 3) The strength of the speech keeps high as a whole

The questioner can't get an appropriate answer for the question and there are many one-sided speech each other in this pattern. Therefore, this shows the gap of the mutual understanding is large. Specifically, the strength of the egocentrism score is 5 or more (Fig. 3).

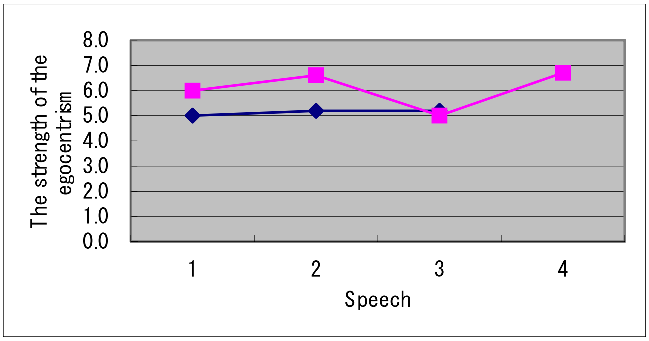


Fig. 3. The example of the strength of the speech keeps high as a whole

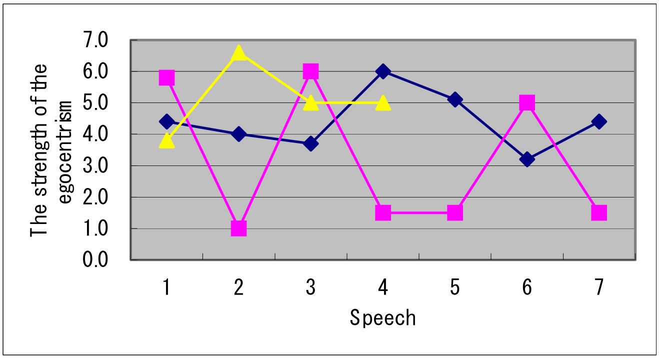


Fig. 4. The example of the other speech

(Pattern 4) The others

The questioner can't get an appropriate answer for the question and there are many one-sided speech, additional information and some proposal in this pattern. Therefore, this shows the gap of the mutual understanding is large (Fig. 4).

Table 2. The examples of the real speech

Classification	Speech ([ ] : Classification of the egocentrism)	Gap
The strength of the last speech is zero.	I am wondering whether the white one or the pink one. Do you know there are any differences 43H and 43H II ? They don't have any differences of the price and functions. [Imposition] Thank you for your answer. [Thanks] It doesn't have any differences of the price and the performance. I think you should choose the color you want, but I recommend the white one. Thank you for your advice. [Thanks]	Small
The strength of just before the last speech is zero.	I have canceled the necessary site by mistake. When I want to use it immediately, what kind of method will there be unless I beg the site? I think that the re-registration is difficult even if asking to your regret when there is an outline that prohibits the agreement for use from registering again. I'm sorry. [Excuse] I think the site will ask you what your purpose is. Do you think so? [Question] Thanks. [Thanks] Since it applies to the agreement, you should wait for it. [Conclusion]	Small
The strength keeps high as a whole.	I think 350 yen for the fully downloaded music is expensive. I can find just music that is not fully downloaded music for 100 yen. Are there any such sites? Because it needs to change fully downloaded music to the sound of calling, it is natural when I have to afford to spend time. [Conclusion] It is legal if you have sources such as CDs, but you need an attention if you don't have them because of illegal action. [Conclusion] I'm sorry, but there are some, which can't be played depending on a model. [Detachment]	Large
Others	When I try to connect to the Web after turning on the power, it shows the initialization screen and disappears all favorite sites. Tell me how to fix it. I think it must be out of order, So, you should go to "au" shop and repair it. [Imposition] The board may need to be changed like me, because it showed initialization and disappeared the favorite sites. [Inference] Finally, you may write only favorite URLs in your e-mail and store in "au My Page" as an e-mail. But I think it is dangerous and can't recommend it because you send it to yourself. [Conclusion]	Large

5 Evaluation Experiment

To verify our method we mentioned above with actual data, we collected text data of 1,585 threads from Japanese Q&A sites. There were 41 threads that speech had the expression of the egocentrism appeared 3 or more times, which were needed to obtain a significant graph shape in the data. We investigated the gap of the mutual understanding with these data and were able to confirm that all threads were same as the results presumed by this method. Table 2 shows the examples of the actual speech and their decisions.

Here, the answer that finally satisfies the demand of the questioner is obtained in the first speech of the examples. Therefore, the egocentrism strength of the last sentence becomes zero and we realize that the gap of the mutual understanding is small as a result. In the second speech of the examples, the egocentrism strength of the last sentence isn't zero, and it is the additional explanation of the previous sentence. Therefore, we realize the gap of the mutual understanding is small. In the third speech of the examples, all speech keeps high state of the egocentrism strength and it can't get final agreement. Concretely, the questioner wants to know cheep full downloaded songs, but the respondent returns that there aren't any sites. Therefore, we can say that the gap of the mutual understanding is large. In the forth speech, the egocentrism strength moves up and down, which is different from the other three patterns, and it has not reached to the final mutual agreement. Concretely, the questioner asks methods to solve the problem, but the respondent replies there is no method for it. Therefore, we realize the gap of the mutual understanding is large.

6 Conclusion

In this paper, we propose the method of presumption for the gap of mutual understanding in dialogues at Q&A sites with the classification and the strength of the egocentrism. It is expected that this method will be applied to monitor the site, because it presumes by the surface of expressions without analyzing deep context of the sentences in the dialogues and high-speed processing is available. In addition, we believe that informing the gap of mutual understanding will be useful in e-mails and general Internet bulletin boards. Furthermore, we picked up the threads that had the expressions of the egocentrism more than three times to get the shape of the strength for the egocentrism in this evaluation. However, it is necessary to examine the presumption method concerning the gap of the mutual understanding for these short speeches because most speech had two or less times of the egocentrism. In the future, we will review by applying more sites and improve the presumption precision. We can also apply this method to select only the information that the gap of the mutual understanding is small from Q&A sites and automatically collect the significant knowledge.

References

1. Joinson, A.: Understanding the Psychology of Internet Behavior. Palgrave Macmillan (2003)
2. Piaget, J.: Piaget's Theory. Carmichael's manual of child psychology 1 (1970)
3. Kruger, J., Eplay, N.: Egocentrism Over E-Mail: Can we Communicate as Well as We Think? Journal of Personality and Social Psychology 89(6), 925–936 (2005)
4. Numa, K., Ohmukai, K., Hamasaki, M., Takeda, H.: Proposal and Implementation of an Egocentric Search Method on Weblog, the Japanese Society of Artificial Intelligence SIG-SWO-A401-06 (2004)
5. Matsuoka, H., Seshimo, H., Okano, S., Arakawa, N., Katou, Y.: Weblog Analysis for Egocentric Communication Support, Technical Report of IEICE KBSE2004-57 (2004)
6. Shimada, S., Fukuhara, T., Satoh, T.: Article Organization Method using Author's Relationships in Question and Answering Bulletin Boards, IEICE DEWS2008 B6-5 (2008)
7. Murata, T., Ikeya, T.: Analysis and Visualization of Internet QA Bulletin Boards Represented as Heterogeneous Networks. Transactions of the Japanese Society of Artificial Intelligence 23(5), 293–302 (2008)
8. Aramaki, E., Abekawa, T., Murakami, Y., Semoto, A.: Detection of Question and Answering Relation between Speech in BBS Dialogues. In: Association for Natural Language Processing NLP 2008, pp. 21–24 (2008)
9. Suzuki, N., Tsuda, K.: Egocentrism Presumption Method with N-gram for e-Business. In: Lovrek, I., Howlett, R.J., Jain, L.C. (eds.) KES 2008, Part II. LNCS (LNAI), vol. 5178, pp. 1002–1009. Springer, Heidelberg (2008)
10. Matsumoto, H.: A morpheme analysis system "ChaSen". Information Processing 41(11) (2000)

Deriving Electrical Dependencies from Circuit Topologies Using Logic Grammar

Takushi Tanaka

Department of Computer Science and Engineering
Fukuoka Institute of Technology
3-30-1 Wajiro-Higashi Higashi-ku, Fukuoka 811-0295, Japan
tanaka@fit.ac.jp
http://www.fit.ac.jp/~tanaka

Abstract. We have developed a new logic grammar for knowledge representation of electronic circuits. The grammar rules not only define the syntactic structure of electronic circuits, but also allow us to derive the meaning of a given circuit as relationships between its syntactic structure and basic circuit functions. In this paper, we show how voltage and current dependencies are coded in this new circuit grammar and how these dependencies are derived by parsing circuit structures.

1 Introduction

When a designed circuit does not work, the engineer tries to localize the fault. He first checks the power supply to confirm the correct voltage is being applied, since the power supply voltage is a prerequisite for correct behavior throughout the whole circuit. Next, he traces causal chains of voltage and current relationships in the circuit. The location of the fault is often determined by finding a place where causal chains fail to connect as intended.

Since inputs and outputs are clearly separated in logic circuits, problems in deriving causal chains do not occur at the logic level. In contrast, it is much harder to derive the causal chains from the circuit topology at the transistor level without knowledge of how circuits are organized. A current through a resistor causes a voltage across that resistor, while, inversely, a voltage applied to a resistor causes a current through the resistor. Although it is difficult to determine which is the cause and which is the effect from the standpoint of the physics of electrical devices, engineers use this kind of causal reasoning to form causal chains that explain how a given circuit works.

One of the first AI studies on deriving causal chains of voltage and current relationships was the work by deKleer[2]. In that work, he tried to derive causal chains from first principles of circuit theory rather than using knowledge of circuit structures. In order to suppress divergence of causal chains, he used heuristics and teleology, and his work led to later AI studies on qualitative reasoning.

We have already developed an approach to circuit structures which can formalize knowledge of how circuits are organized. This work was based on the

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 325–332, 2009.

idea that electronic circuits are designed as hierarchical structures of functional blocks. Since these hierarchical structures were analogous to the syntactic structures of language, we were able to develop a grammatical method for parsing electronic circuits[4]. In that study, we viewed each circuit as a sentence and its elements as words. Structures of functional blocks were defined by a logic grammar called DCSG (Definite Clause Set Grammar)[3]. DCSG was successful in representing knowledge of circuit structures. This approach, however, was not able to represent knowledge of circuit functions.

The newly developed circuit grammar has fields for semantic terms. Using these semantic terms, we can define relationships between circuit structures and their functions. Here, we assume that circuit functions are the meaning of the circuit structures. The circuit functions we consider are the electrical behaviors that are useful to circuit designers or users. These electrical behaviors are defined on the voltages and currents occurring in the circuit. In particular, electrical dependencies such as causality and conditions are useful to understand how circuits work. In this paper, we show how voltage and current dependencies are coded in the new circuit grammar, and how these dependencies are derived through parsing circuit structures.

2 Circuit Grammar with Semantic Term

The new circuit grammar is an extension to DCSG (Definite Clause Set Grammar) [3] which is a DCG[1]-like logic grammar developed for analyzing word-order free languages. The new circuit grammar can define relationships between circuit structures and functions using semantic terms in grammar rules.

2.1 Semantic Terms in the Left-Hand Side

Semantic terms are placed in curly brackets in grammar rules as follows.

$$A, \{F_1, F_2, ..., F_m\} \ \longrightarrow \ B_1, B_2, ..., B_n. \tag{1}$$

This grammar rule can be read as stating that the symbol A with meaning $\{F_1, F_2, ..., F_m\}$ consists of the syntactic structure $B_1, B_2, ..., B_n$. This rule is converted into a Prolog clause as follows.

$$\begin{aligned} &ss(A, S_0, S_n, E_0, [F_1, F_2, ..., F_m|E_n]) :- \\ &\qquad\qquad ss(B_1, S_0, S_1, E_0, E_1), \\ &\qquad\qquad ss(B_2, S_1, S_2, E_1, E_2), \\ &\qquad\qquad \cdots\ , \\ &\qquad\qquad ss(B_n, S_{n-1}, S_n, E_{n-1}, E_n). \end{aligned} \tag{1'}$$

When a rule is used in parsing, the goal $ss(A, S_0, S_n, E_0, E)$ is executed, where the variable S_0 is replaced by an object set (object circuit) and the variable E_0 is replaced by an empty set. The subsets (sub-circuits) "$B_1, B_2, ..., B_n$" are successively identified in the object set S_0. After all of these subsets are identified, the remainder of these subsets (the complementary set) is put into S_n. While, the semantic information of B_1 is added with E_0 and put into E_1, the semantic information of B_2 is added with E_1 and put into E_2,..., and the semantic information of B_n is added with E_{n-1} and put into E_n. Finally, the semantic information $\{F_1, F_2, ..., F_m\}$, which is the meaning associated with symbol A, is added and all of the semantic information is put into E.

Terminal symbols are surrounded by square brackets in grammar rules. The symbol "$[B_i]$" is converted to $member(B_i, S_i, S_{i+1})$ which identifies the element B_i in the object set S_i, and put the remainder into S_{i+1}. The terminal symbol "$[B_i]$" does not change the current semantic information E_i, but the technique in Section 3.2 enables us to add semantic informations to terminal symbols. Here, the predicate member is defined as:

$$member(M, [M|X], X). \tag{2}$$
$$member(M, [A|X], [A|Y]) :- \ member(M, X, Y). \tag{3}$$

2.2 Semantic Terms in the Right-Hand Side

Semantic terms in the right-hand side define the semantic conditions for the grammar rule. For example, the following rule (4) is converted into the Prolog clause (4)' as follows.

$$A \longrightarrow B_1, \{C_1, C_2\}, B_2. \tag{4}$$

$$\begin{aligned} ss(A, S_0, S_n, E_0, E_n) \ :- \ & ss(B_1, S_0, S_1, E_0, E_1), \\ & member(C_1, E_1, \_\ ), \\ & member(C_2, E_1, \_\ ), \\ & ss(B_2, S_1, S_2, E_1, E_2). \end{aligned} \qquad (4)'$$

When the clause (4)' is used in parsing, the conditions C_1 and C_2 are tested to see if the semantic information E_1 fills these conditions after identifying the symbol B_1. If it succeeds, the parsing process goes on to identify the symbol B_2.

3 Coding Electrical Dependencies

3.1 Circuit Representation

We now develop grammar rules using the functional blocks appearing in the circuit $cd15$, which is a type of operational amplifier called a transconductance amplifier (Figure 1). The circuit $cd15$ is represented as the following word-order free sentence. Here, the compound term $npnTr(q1, 3, 5, 6)$ is a terminal symbol which represents the NPN-transistor named $q1$ with the base connected to node 3, the emitter to node 5, and the collector to node 6 respectively.

$$\begin{aligned} \mathrm{cd15}([\ & npnTr(q1, 3, 5, 6), npnTr(q2, 4, 5, 7), npnTr(q3, 10, 1, 5), \\ & npnTr(q4, 10, 1, 10), npnTr(q5, 8, 1, 8), npnTr(q6, 8, 1, 9), \\ & pnpTr(q7, 6, 2, 6), pnpTr(q8, 6, 2, 9), pnpTr(q9, 7, 2, 7), \\ & pnpTr(q10, 7, 2, 8)]). \end{aligned} \tag{5}$$

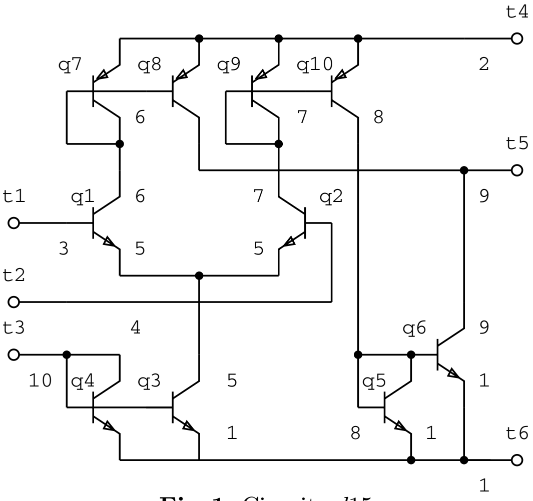


Fig. 1. Circuit *cd*15

3.2 Rules for Elements and Devices

The grammar rule (6) defines an NPN-transistor Q in active state. Although "$npnTr(Q, B, E, C)$" is a terminal symbol, it is also defined as a non-terminal with semantic information. The compound term $gt(v(C, E), vst)$ represents the fact that the collector-emitter voltage $v(C, E)$ is greater than the collector saturation voltage vst. The compound term $equ(v(B, E), vbe)$ represents the fact that the base-emitter voltage $v(B, E)$ is equal to the forward voltage of p-n junction vbe. The compound term $cause(v(B, E), i(B, Q), Q)$ represents the fact that the base-emeitter voltage $v(B, E)$ causes the base current $i(B, Q)$ by the operation of the NPN-transistor Q. Here, $i(B, Q)$ represents the branch current from node B to transistor Q.

Grammar rules for the saturated state and the cutoff state are also defined. When a terminal symbol $[npnTr(Q, B, E, C)]$ is found in parsing a circuit, one of these rules is selected, and its semantic terms are derived as a meaning of the symbol non-deterministically. Similar rules are also defined for PNP-transistors.

$$\begin{aligned}
&npnTr(Q, B, E, C),\\
&\{\ state(Q, active),\\
&\quad gt(v(C, E), vst),\\
&\quad equ(v(B, E), vbe),\\
&\quad gt(i(B, Q), 0),\\
&\quad gt(i(C, Q), 0),\\
&\quad cause(v(B, E), i(B, Q), Q),\\
&\quad cause(v(B, E), i(Q, E), Q),\\
&\quad cause(i(B, Q), v(B, E), Q),\\
&\quad cause(i(Q, E), v(B, E), Q),\\
&\quad cause(i(B, Q), i(C, Q), Q)\} \ \longrightarrow\ [npnTr(Q, B, E, C)].
\end{aligned} \tag{6}$$

3.3 Rules for Functional Blocks

A transistor in which the base and the collector are connected together works as a diode (Figure 2). The following grammar rule (7) defines the diode-connected transistor "$dtr(dtr(Q), A, C)$" in the *conductive* state as a non-terminal symbol. The syntactic part of the right-hand side defines either an NPN-transistor Q or a PNP-transistor Q whose base and collector are connected to the same node. The semantic term $state(Q, active)$ in the right-hand side is an electrical condition which requires that the transistor Q must be in the *active* state. Here, $dtr(Q)$ is a name given to the diode (Skolem function).

The semantic terms in the left-hand side represent the electrical behavior in the conductive state. E.g. "$gt(i(A, dtr(Q)), 0)$" represents the current flow from A to $dtr(Q)$. Here, "$i(A, dtr(Q))$" is the branch current from the node A to the functional block $dtr(Q)$. Since we assume that branch current flows not only from a node to an element but also from a node to a functional block, the same current can have different expressions, for example $i(A, dtr(Q))$ and $i(A, Q)$. The semantic term $equiv(i(dtr(Q), C), i(Q, A))$ is added to form the equivalence relations for these currents. The grammar rule for diode-connected transistor in reverse bias is also defined in the same manner.

$$\begin{aligned}
&dtr(dtr(Q), A, C),\\
&\{\ state(dtr(Q), conductive),\\
&\quad gt(i(A, dtr(Q)), 0),\\
&\quad cause(v(A, C), i(A, dtr(Q)), dtr(Q)),\\
&\quad cause(v(A, C), i(dtr(Q), C), dtr(Q)),\\
&\quad cause(i(A, dtr(Q)), v(A, C), dtr(Q)),\\
&\quad cause(i(dtr(Q), C), v(A, C), dtr(Q)),\\
&\quad equiv(i(A, dtr(Q)), i(A, Q)),\\
&\quad equiv(i(dtr(Q), C), i(Q, C))\} \longrightarrow\\
&\qquad\qquad\qquad\qquad (npnTr(Q, A, C, A);\ pnpTr(Q, C, A, C)),\\
&\qquad\qquad\qquad\qquad \{state(Q, active)\}.
\end{aligned} \tag{7}$$

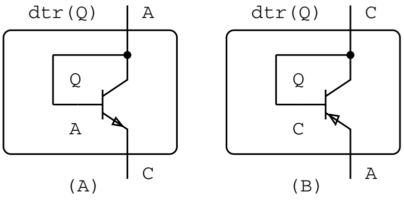


Fig. 2. Diode-connected transistor

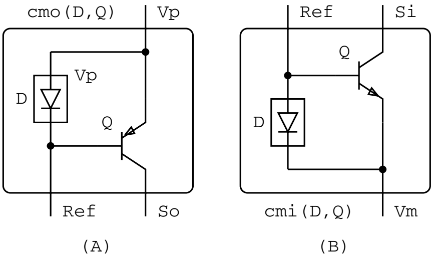


Fig. 3. Current mirror

Figure 3 shows two current mirror circuits. Both circuits generate the same current as their reference current. The grammar rule (8) is defined for the source-type current mirror shown in Figure 3(A). The semantic terms in the right-hand side are the electrical conditions that operate the circuit. The semantic term in the left-hand side "$cause(i(cmo(D,Q),Ref),i(cmo(D,Q),So),cmo(D,Q))$" is related to the main function of this circuit. That is, the external current from $cmo(D,Q)$ to Ref causes another external current from $cmo(D,Q)$ to So by the circuit $cmo(D,Q)$. The next "$cause(i(cmo(D,Q),Ref),i(D,Ref),cmo(D,Q))$" represents the external current $i(cmo(D,Q),Ref)$ causes the internal current $i(D,Ref)$ of the functional block. This causal relationship connecting external and internal aspects of functional block enables us to explain the main function of the functional block togather with equivalence relations between currents.

$$\begin{aligned}
&currentMirrorSource(cmo(D,Q),Ref,Vp,So),\\
&\{\ cause(i(cmo(D,Q),Ref),i(cmo(D,Q),So),cmo(D,Q)),\\
&\quad cause(i(cmo(D,Q),Ref),i(D,Ref),cmo(D,Q)),\\
&\quad equiv(i(cmo(D,Q),So),i(Q,So))\} \longrightarrow\\
&\qquad\qquad\qquad dtr(D,Vp,Ref),\\
&\qquad\qquad\qquad \{state(D,conductive)\},\\
&\qquad\qquad\qquad pnpTr(Q,Ref,Vp,So),\\
&\qquad\qquad\qquad \{state(Q,active)\}.
\end{aligned} \tag{8}$$

Grammar rules for the emitter-coupled pair (Figure 4) and the transconductance amplifier (Figure 5) are also defined in the same manner.

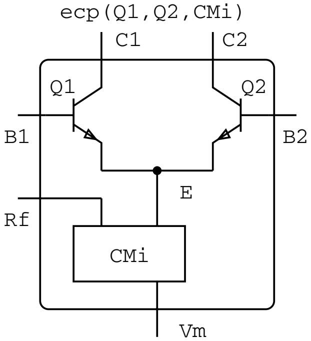


Fig. 4. Emitter coupled pair

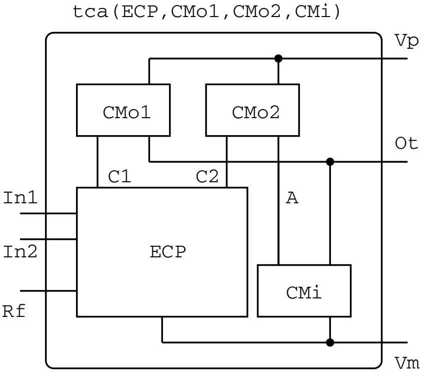


Fig. 5. Transconductance amp

4 Parsing Circuits

All of the grammar rules defined in the previous section are converted into Prolog clauses according to the circuit grammar conversion method described in Section 2. The clauses form a logic program that performs top-down parsing. The following goal (9) parses the circuit $cd15$ and derives the circuit structure and its electrical behaviour. The first subgoal $cd15(CT)$ substitutes the circuit $cd15$ into the variable CT. The circuit is given to the second argument of the predicate $ss(...)$. The first argument X is a functional block identified in the circuit and the third argument is the remainder of the circuit. Since the third

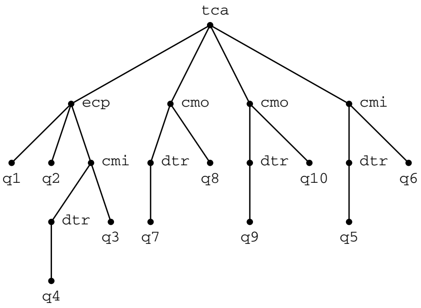


Fig. 6. Parse tree for *cd*15

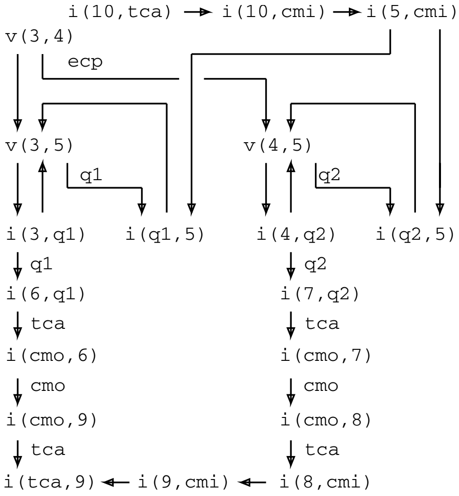


Fig. 7. Causal Chains on *cd*15

argument is empty, the goal asks whether the whole circuit can be identified as the single non-terminal symbol X. The fourth argument [ ] means no semantic information is given at the start of parsing. Each time a functional block is identified, semantic information about the functional block is added. After the whole circuit is parsed, the value of Y has much semantic information about the circuit.

$$?-\ cd15(CT),\ ss(X,CT,[\,],[\,],Y). \tag{9}$$

$$\begin{aligned} X =\ & transCondAmp(\ tca(\ ecp(q1,q2,cmi(drt(q4),q3)),\\ &\qquad\qquad\qquad\qquad cmo(dtr(q7),q8),\\ &\qquad\qquad\qquad\qquad cmo(dtr(q9),q10),\\ &\qquad\qquad\qquad\qquad cm1(dtr(q5),q6)),\\ &\qquad\qquad\qquad 3,4,10,2,9,1) \end{aligned}$$

$$\begin{aligned} Y =\ & [\,cause(v(3,4),i(tca(.),9),tca(.)),\\ & \ cause(i(cmo(.),9),i(tca(.),9),tca(.)),\\ & \ cause(i(9,cmi(.)),i(9,tca(.)),tca(.)),\\ & \ equiv(i(10,tca(.)),i(10,ecp(.))),\\ & \ \ ...\ 187\ lines\ omitted\ ...)] \end{aligned}$$

The value of X shows that the circuit *cd*15 is identified to be the operational amplifier "*transCondAmp*(*tca*(...), 3, 4, 10, 2, 9, 1)". The first argument *tca*(...) is a name given to the identified circuit, and the rest are the connecting nodes in the circuit. The name keeps track of identified functional blocks and is viewed as a parse tree which shows the syntactic structure of the circuit (Figure 6). Each node represents a functional block identified in the circuit *cd*15.

The semantic information substituted into Y consists of electrical conditions and causal relationships. The electrical conditions show the conditions under which all the functional blocks will work correctly as components of the given circuit. The causal relationships consist of dependencies on voltages and currents. The first one "$cause(v(3,4), i(tca(.), 9), tca(.))$" is added after identifying the whole circuit. It is related to the main function of transconductance amplifier. Here, the structure of the transconductance amplifier is abbreviated as $tca(.)$. This main causal relationship is also supported by internal causal chains of the transconductance amplifier as shown in Figure 7. The causal relationships such as from $i(cmo, 7)$ to $i(cmo, 8)$ in the figure are also supported by internal causal chains of source-type current mirror circuits. Causal relationships and equivalence relations (Section 3.3), which connect internal and external expressions of branch currents, enable these supports.

5 Conclusions

This newly developed circuit grammar has fields for semantic terms. Using these semantic terms, we have defined electrical conditions to allow the circuit functions and causal relationships between voltage and current to be used to explain the circuit's behavior. In particular, we derive these electrical dependencies as the meanings of the circuit structures by parsing syntactic structures. These derived informations will be useful both for understanding a circuit's structure and behavior and for troubleshooting.

The derived semantic terms can also be viewed as a word-order free sentence. Here, "$cause(...)$" and "$equiv(...)$" are terminal symbols. We can easily define non-terminal symbols which implement transitivity on causality and equivalence relations on branch currents[5]. These electrical dependencies, however, only describe the shallow behaviour of electrical circuits. We are currently developing a language for describing circuit behaviours and functions more precisely.

References

1. Pereira, F.C.N., Warren, D.H.D.: Definite Clause Grammars for Language Analysis. Artificial Intell. 13, 231–278 (1980)
2. de Kleer, J.: Causal and Teleological Reasoning in Circuit Recognition. MIT, AI-TR-529 (1979)
3. Tanaka, T.: Definite Clause Set Grammars: A Formalism for Problem Solving. J. Logic Programming 10, 1–17 (1991)
4. Tanaka, T.: Parsing Circuit Topology in A Logic Grammar. IEEE-Trans. Knowledge and Data Eng. 5, 225–239 (1993)
5. Tanaka, T.: A Logic Grammar for Circuit Analysis - Problems of Recursive Definition. In: Apolloni, B., et al. (eds.) KES 2007, Part II. LNCS (LNAI), vol. 4693, pp. 852–860. Springer, Heidelberg (2007)

A Fast Nearest Neighbor Method Using Empirical Marginal Distribution

Mineichi Kudo, Jun Toyama, and Hideyuki Imai

Division of Computer Science
Graduate School of Information Science and Technology
Hokkaido University
Kita-13, Nishi-9, Kita-ku, Sapporo 060-0814, Japan
{mine,jun,imai}@main.ist.hokudai.ac.jp
http://prml.main.eng.hokudai.ac.jp

Abstract. Unfortunately there is no essentially faster algorithm than the brute-force algorithm for the nearest neighbor searching in high-dimensional space. The most promising way is to find an approximate nearest neighbor in high probability. This paper describes a novel algorithm that is practically faster than most of previous algorithms. Indeed, it runs in a sublinear order of the data size.

1 Introduction

The k-nearest neighbor (k-NN) method is widely used in pattern recognition. Since, for each query, the brute-force algorithm needs $O(nm)$ time in a dataset of m-dimensional n samples, many studies have been devoted for speeding up the query time. However, as m grows, say over 100, almost all sophisticated algorithms become slower than the brute-force algorithm (the *curse of dimensionality*). The only one solution to cope with the curse of dimensionality seems to introduce some kind of tolerance into the problem. Typically, there are two kinds of such tolerance. The first one is "approximation." With this kind of tolerance, we are satisfied with suboptimal k nearest neighbors in the sense that the solution has a distance less than $(1+\eta)$ times the distance to the true kth nearest neighbor. Such a trial is called *approximate nearest neighbors* (ANN framework) [1,2]. The ANN algorithm in [2] requires $O(mn \log n)$ time and $O(mn)$ space for preprocessing and runs in $O(c_{m,\eta} \log n)$ with $c_{m,\eta} \leq m\lceil 1 + 6m/\eta \rceil^m$. Note that the upperbound grows exponentially in m in the worst case. Kleinberg [3] also proposed an algorithm that has a near-linear storage and query time in m when the approximate nearest neighbor is the goal to attain.

The other kind of tolerance is "probabilistic correctness." With this kind of tolerance, the complete correctness of a solution is not guaranteed, but the error probability is upperbounded. In this case, a *confidence (error)* parameter ϵ is introduced to upperbound the error probability of the algorithm missing the true k nearest neighbors. Such a trial is called *probably correct nearest neighbors* (PCNN framework). A study in this line is shown in [4]. Both kinds of tolerance

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 333–339, 2009.

can be considered simultaneously. Then we come to *probably and approximately correct nearest neighbors* (PACNN framework). Such a trial is introduced in [5], although only a naive algorithm is shown in a standard framework. Probably the fastest algorithms for PACNN are locality-sensitive hashing algorithms [6,7,8]. They adopt multiple sets of hashing functions randomly chosen for narrowing the search area and show sublinear algorithms in n.

In this study, the PCNN framework will be considered. At the expense of ϵ error in the correctness, we can reduce the searching time at the rate δ. This does not change the time complexity, but the algorithm runs in $O((1-\delta)nm)$ for data size n and dimensionality m. That is, it runs necessarily faster than the brute-force algorithm whose order is $O(nm)$. There are some ANN algorithms that run in $O(\log n)$ and some PACNN algorithms run in $O(mn^c)$ with $c < 1$ in theory. Therefore our algorithm appears to be slower than such algorithms, but in practice it can be faster than them. This is because the cost necessary for search processing has a great deal of effect on real searching time. For example, in the algorithm in [2], the coefficient grows exponentially in m. In our algorithms, however, there is no factor affected by m. The effectiveness depends on only the nature of the dataset, so that easier problems are solved much faster in the searching.

2 Key Idea

Simply speaking, our algorithm relies on the fact that "a very close pair in the original m dimensions is also close in the first few l dimensions in high probability." Then we can exempt a sample from full-distance calculation if it is far from the query point in the l dimensions. It would bring a large degree of efficiency when l is far less than m. In the following, we describe this strategy formally for large m and small but more than one l. The point is that l is fairly less than m, so that l can be regarded as a constant.

For simplicity, we consider the case of $k = 1$, that is, the nearest neighbor search. Indeed, when the dimensionality m is large, it can be shown theoretically that the computation cost is the same for any value of k as long as k is not so large. We firstly notice that the ranking of points changes if some coordinates are dropped and that 1-NN changes as well. Therefore, for probabilistic evaluation, we have to analyze how the following two probabilities change with a threshold θ_l: $P(D_l^2(X, X_{1NN}) < \theta_l)$ of the squared distance between a point X and its nearest neighbor X_{1NN} and $P(D_l^2(X, Y) < \theta_l)$ of the squared distance between X and any other point Y. Here, $D_l^2(X, Y)$ is the squared Euclidean distance measured in the first l dimensions, that is, $D_l^2(X, Y) = \sum_{j=1}^{l}(X_j - Y_j)^2$ for $X = (X_1, X_2, \ldots, X_m)$ and $Y = (Y_1, Y_2, \ldots, Y_m)$.

When we adopt $\theta_l = D_m^2(X, \hat{X}_{1NN})$ regardless of the value of l, where $\hat{X}_{1NN}$ is the current candidate of the nearest neighbor in the middle of the search, then we come to the *partial distance* algorithm. In this case, in the stored sample set, $P(D_l^2(X, X_{1NN}) < \theta_l) = 1$ because $D_l^2(X, X_{1NN}) \leq D_m^2(X, X_{1NN}) \leq D_m^2(X, \hat{X}_{1NN}) = \theta_l$ holds. In this paper, we consider its probabilistic version.

In this version, we establish the value of θ_l so as to make the *misjudgement error* $\epsilon = P(D_l^2(X, X_{1NN}) > \theta_l)$ sufficiently small and to make *reduction probability* $\delta = P(D_l^2(X, Y) > \theta_l)$ as large as possible. Note that $D_l^2(X, X_{1NN}) \leq D_m^2(X, X_{1NN})$ always holds for $l \leq m$, so that we can expect a smaller value of θ_l than $D_m^2(X, \hat{X}_{1NN})$.

3 Algorithm

Our *marginal distance strategy* (*MDS*) algorithm is very simple. First we find the best order of coordinates using principal component analysis. Next, we obtain the empirical distribution functions $\hat{F}(D_l^2(X, X_{1NN}))$ and $\hat{G}(D_l^2(X, Y))$ for $l = 1, 2, \ldots, l_{max}$ in the principal coordinates. Then for a specified value of ϵ, we determine the value of θ_l as the minimum value satisfying $1 - \hat{F}(\theta_l) < \epsilon$ (*empirical percentile approach*). At the same time, we can know the estimate of δ as $1 - \hat{G}(\theta_l)$. In the searching phase, we exempt sample Y from the full-dimensional search if $D_l^2(X, Y) > \theta_l$ for some l. Here, $\hat{F}$ and $\hat{G}$ are both estimated from the samples.

3.1 Preprocessing

First, for a prespecified value of l_{max}, we find l_{max} eigen vectors of the covariance matrix $\hat{\Sigma}$ estimated from all of the training samples. Next, we randomly choose n' samples (n' fixed at 1000 in the following experiment) from all n samples. For these n' samples, we find their kth nearest neighbors. We then project all of the training samples into the l_{max}-dimensional subspace spanned by the l_{max} principal (eigen) vectors. In each l-dimension ($l = 1, 2, \ldots, l_{max}$), we obtain two empirical densities of $D_l^2(X, X_{kNN})$ and $D_l^2(X, Y)$, where X and Y are taken only from n' samples, but X_{kNN} is found from all n samples. In this way, we obtain two empirical distributions $\hat{F}_l$ and $\hat{G}_l$.

3.2 User Choice of Parameters

After we obtain $\hat{F}_l$ and $\hat{G}_l$ ($l = 1, 2, \ldots, l_{max}$), we prompt the user to specify the value of confidence (error) parameter ϵ and the value of marginal dimension l. For helping the user, two estimated reduction ratios are presented for several candidate values of ϵ and l: the expected reduction rate δ_l of full distance calculation and the expected reduction rate δ_l^* of searching time. Here, δ_l is directly obtained from the empirical distribution $1 - \hat{G}_l(D_l^2(X, Y) = \theta_l)$, where θ_l is taken so as to satisfy $\hat{F}_l(D_l^2(X, X_{kNN}) = \theta_l) \simeq 1 - \epsilon$, while δ_l^* is calculated from δ_l by Eq. (1) which will be given later. For the value of l, the user is allowed to use the estimated optimal value l_{opt} which will also be given later in Eq.(2). So, the user may choose only the value of ϵ. Then the user knows the values of δ_l and δ_l^* before searching.

```
0.      Preparation:
          $q \leftarrow$ query point
          $q^p \leftarrow \Phi^t q$    /* projection to the $l$-dimensional space */
          $r_k^* \leftarrow +\infty$
          $1NN, \ldots, kNN \leftarrow 0$

        Search:
1.s     for $i = 1$ to $n$
1.1.s     $r^2 \leftarrow 0$
          for $j = 1$ to $l$
            $r^2 \leftarrow r^2 + (x_{ij}^p - q_j^p)^2$
          endfor
          if $(r^2 > \theta_l)$      /* marginal distance thresholding */
            Break the process of the $i$th sample and proceed to $(i+1)$th
1.1.e     endif
1.2.s     $r^2 \leftarrow 0$
          for $j = 1$ to $m$
            $r^2 \leftarrow r^2 + (x_{ij} - q_j)^2$
            if $(r^2 > (r_k^*)^2)$      /* partial distance thresholding */
              Break the process of the $i$th sample and proceed to $(i+1)$th
            endif
1.2.e     endfor
1.3.s     if $(r^2 < (r_k^*)^2)$       /* update of records */
            Update $1NN, \ldots, kNN$ with distances $r_1, r_2, \ldots, r_k$
            $(r_k^*)^2 \leftarrow r_k^2$       /* update of the $k$th NN distance */
1.3.e     endif
1.e     endfor
2.      output $1NN, \ldots, kNN$
```

Fig. 1. Searching algorithm MDS

3.3 Searching Algorithm

The searching algorithm is shown in Fig.1. Steps 1.1.s-e show the marginal strategy using the projected l-dimensional vectors. Apparently a continuous judgement in $j = 1, 2, \ldots, l$ is more effective, but we judge only at $j = l$. This is for making the actual error close to the estimation. In Steps 1.2.s-e, we restart the distance calculation from the first dimension using the original vectors instead of the projected vectors. We do this because the cost of obtaining all of the m projected values of the query point is higher than that of this way. Steps 1.3.s-e are the ordinal update procedure of the current solution.

It should be noted that the MDS strategy (Step 1.1) is used as a filter for data screening and is independent of the following procedures. Indeed, Steps 1.1 and 1.2 are independent to each other. Therefore, Step 1.2 can be replaced with another sophisticated algorithm.

3.4 Analysis of Algorithm and the Optimal Marginal Dimension

Let us examine the space complexity necessary both for preprocessing and for searching. When we use l dimensions for MDS, in addition to keeping all n data, we need $O(ln)$ for storing the projected n vectors. In addition, we need $O(lm)$ for keeping l principal (eigen) vectors of length m. Here, l is usually small enough. As a result, the memory storage is still $O(nm)$, which is the minimum requirement for keeping all n training data. For $n \gg m$, it can be estimated as $O(n)$.

Next, let us clarify the time complexity. In the preprocessing, we have to obtain l principal vectors from the estimated covariance matrix $\hat{\Sigma}$, Roughly $O(nm^2)$ is needed for calculation of $\hat{\Sigma}$ and $O(ml)$ is needed for calculation of l principal vectors. We estimate two distributions $F_l(D_l^2(X, X_{kNN}))$ and $G_l(D_l^2(X,Y))$ $(l = 1, 2, \ldots, l_{max})$ from $n'(< n)$ sampled points. We need $O(n'n)$ for the former distribution and $O(n'^2)$ for the latter distribution. Thus, we can regard it as $O(n)$ for $m, n' \ll n$.

For analyzing the actual query time, let us assume that cost c_1 is consumed for a one-dimensional distance operation (one subtraction, one multiplication, and one addition) and c_2 for the inner-product operation (one multiplication plus one addition). Then we can estimate the query time T_l as

$$\begin{aligned} T_l &= (1 - \delta_l^*)T_0, \\ &\text{where} \quad T_0 = c_1 nm, \quad \delta_l^* = \delta_l - \frac{c_2 l}{c_1 n} - \frac{l}{m}. \end{aligned} \tag{1}$$

Here, T_0 is the cost consumed in the brute-force algorithm and δ_l^* is the estimated time reduction rate (δ_l being the ratio at which full distance calculation is skipped). Only when $\delta_l > l/m$, it is possible to reduce the searching time.

For optimal setting of the value of marginal dimension l, we can use Eq. (1) as

$$l_{opt} = \arg\max_l \delta_l^* \left(= \delta_l - \frac{c_2 l}{c_1 n} - \frac{l}{m} \right). \tag{2}$$

Here, $\delta_l = \hat{G}_l(D_l^2(X,Y) = \theta_l)$ and θ_l is determined by specified ϵ. In the following experiments, we determined the values of c_1 and c_2 experimentally.

4 Experiments

We used the MNIST dataset [9]. It contains 60 000 points, each having dimension $m = 784 (= 28 \times 28)$. It provides another test set of 10 000 points. According to [7], the points were normalized so that each point has its norm equal to one.

We compared the MDS algorithm with the ANN algorithm [2][1] and E^{2}LSH[2] [7]. The ANN algorithm belongs to ANN framework and the E^{2}LSH belongs to PACNN framework. On the contrary, MDS belongs to PCNN framework. The ANN algorithm implements a k-d tree structure tuned for ANN. It also employs an incremental distance update technique.

[1] The code is available at http://www.cs.umd.edu/~mount/ANN/.

[2] The code is available at http://web.mit.edu/andoni/www/LSH/.

Table 1. Comparison of ANN, E^{2}LSH and MDS in the MNIST dataset with $n = 60000$ and $m = 784$. The recognition rate of the correct 1-NN is 87.55% and the searching time of the brute-force algorithm is 2565 seconds.

	ANN			E^{2}LSH			MDS		
Closeness $1+\eta$	1.0	5.0	6.0	—	—	—	1.0	1.0	1.0
Radius R	—	—	—	0.78	0.70	0.65	—	—	—
Confidence $1-\epsilon$	1.0	1.0	1.0	0.90	0.90	0.95	0.999	0.99	0.90
Precision of 1NN(%)	100.00	98.91	97.97	99.81	98.82	96.79	99.96	99.51	97.12
Recognition Rate (%)	87.55	87.34	87.37	87.41	86.64	85.07	87.57	87.55	87.35
Searching Time (s)	1144.3	187.7	130.6	581.8	319.3	270.0	75.1	56.8	52.9
Actual $1+\eta$ (ave)	1.000	1.004	1.009	1.000	1.000	1.000	1.000	1.001	1.009
$1+\eta$ (max)	1.000	1.462	1.623	1.044	1.047	1.019	1.034	1.175	1.740

The ANN has an *approximation parameter* η to guarantee $D(X, \hat{X}_{1NN}) \leq (1+\eta)D(X, X_{1NN})$ (where, $\hat{X}_{1NN}$ is the resultant candidate for 1NN obtained by the algorithm), and E^{2}LSH has a *radius parameter* R and a *confidence (error) parameter* ϵ to guarantee $P(D(X, \hat{X}_{1NN}) \leq R) \geq 1-\epsilon$. MDS has a confidence parameter ϵ to guarantee $P(D(X, \hat{X}_{1NN}) = D(X, X_{1NN})) \geq 1-\epsilon$. In E^{2}LSH, we used $R = 0.65, 0.70$ and 0.78. The other parameters were set to several values.

The results are shown in Table 1. Although it is difficult to fairly compare algorithms belonging to different frameworks, the probability of correct nearest neighbors $P(D(X, \hat{X}_{1NN}) = D(X, X_{1NN}))$ can be used as a common performance indicator. After testing several values of parameters, we obtained some comparable results. We can see that MDS is superior to the other two in searching time for comparable probabilities of correct nearest neighbors.

It should be noted that we cannot know how many nearest neighbors are correct in the ANN approach, although all the answers are close to the correct nearest neighbors up to $(1+\eta)$ times. On the other hand, in the PCNN approach, we cannot know how far the obtained nearest neighbor is from the correct nearest neighbor when the answer is not correct. In the PACNN approach, both are unknown. In these respects, they might be largely different from the correct nearest neighbor searching. However, such anxiety was not recognized in the MNIST dataset. In ANN, the precision of the correct nearest neighbors is sufficiently high even for relatively large η given in the algorithm. On the other hand, in MDS, the actual approximation ratio $1+\eta$ was close to one.

We have also confirmed that MDS algorithm is faster than a PACNN algorithm [5] in many cases in a comparable setting.

5 Discussion

The marginal strategy is generally more effective in a classification task than in a simple searching task without class labels. This is because in classification problems we can expect that the nearest neighbors are mostly generated according to the density function of the class that the query sample belongs to. In general, $\Sigma_T = \Sigma_W + \Sigma_B$, where Σ_T is the total covariance matrix, Σ_W is the within-class covariance matrix, and Σ_B is the between-class covariance matrix. It is clear that $\Sigma_T > \Sigma_W$ (in element-wise). For a query sample X, its k-nearest neighbors can be expected to being generated from the same class as X, and their distance is therefore estimated from Σ_W, while for an arbitrary sample Y, the distance between X and Y is estimated from Σ_T. Therefore, we can expect a large difference in the two distributions of $G(D^2(X, Y))$ and $F(D^2(X, X_{kNN}))$. As a result, we can expect a large reduction value of δ even for a small error value of ϵ.

6 Conclusion

We have proposed a simple k-nearest neighbor search algorithm that works well for high dimensional cases. This algorithm is based on the empirical marginal distributions and is advantageous compared to other algorithms only when data size and dimensionality are both large. The algorithm looses the correctness of the nearest neighbors, but the loss of correctness can be controlled to be arbitrarily small by the user.

References

1. Nene, S.A., Nayar, S.K.: A Simple Algorithm for Nearest Neighbor Search in High Dimensions. IEEE Transactions on Pattern Analysis and Machine Intelligence 19, 989–1003 (1997)
2. Arya, S., et al.: An optimal algorithm for approximate nearest neighbor searching fixed dimensions. Journal of the ACM 45-6, 891–923 (1998), `http://www.cs.umd.edu/~mount/ANN/`
3. Kleinberg, J.M.: Two algorithms for nearest-neighbor search in high dimension. In: Proc. 29th Annu. ACM sympos. Theory Comput., pp. 599–608 (1997)
4. Maneewongvatana, S., Mount, D.M.: An Empirical Study of a New Approach to Nearest Neighbor Searching. In: Deng, R.H., Qing, S., Bao, F., Zhou, J. (eds.) ICICS 2002. LNCS, vol. 2513, pp. 172–187. Springer, Heidelberg (2002)
5. Ciaccia, P., Patella, M.: PAC nearest neighbor queries: Approximate and controlled search in high-dimensional and metric spaces. In: Proceedings of the 16th International Conference on Data Engineering, pp. 244–255 (2000)
6. Indyk, P., Motwani, R.: Approximate nearest neighbors: towards removing the curse of dimensionality. In: Proceedings of the 30th Annual ACM Symposium on Theory of Computing, pp. 604–613 (1998)
7. Andoni, A., et al.: Locality-Sensitive Hashing Using Stable Distributions. In: Shakhnarovich, G., Darrell, T., Indyk, P. (eds.) Nearest-Neighbor Methods in Learning and Vision: Theory and Practice, vol. 3. MIT Press, Cambridge (2006)
8. Andoni, A., Indyk, P.: Near-Optimal Hashing Algorithms for Approximate Nearest Neighbor in High Dimensions. Communications of the ACM 51(1), 117–122 (2008)
9. Le Cunn, Y.: The mnist dataset of handwritten digits, `http://yann.lecun.com/exdb/mnist/`

Effects of Kurtosis for the Error Rate Estimators Using Resampling Methods in Two Class Discrimination

Kozo Yamada, Hirohito Sakurai, Hideyuki Imai, and Yoshiharu Sato

Graduate School of Information Science and Technology, Hokkaido University
Kita 14, Nishi 9, Kita-ku, Sapporo, 060-0814, Japan
{yamada,sakurai,imai,ysato}@main.ist.hokudai.ac.jp

Abstract. In preceding studies, error rate estimators have been compared under various conditions and in most cases the population distribution was assumed to be normal. Effects of non-normality of the population have therefore not been studied sufficiently. In this study, we focused on kurtosis as a measure of non-normality and examined the effects of kurtosis for error rate estimators, especially resampling-based estimators. Our simulation results in two-class discrimination using a linear discriminant function suggest that it is necessary to consider non-normality of the population in comparison of estimators.

Keywords: actual error rate, leave-one-out, bootstrap, smoothed bootstrap, bias, standard deviation, robustness.

1 Introduction

The most widely used criterion for evaluating the performance of a discrimination rule is error rate, namely, the proportion of future observations that are misclassified. Many error rate estimators have been proposed. Much interest has been shown in estimators based on a resampling method since they can be applied to any population distribution. Resampling methods include *cross-validation*, *jackknife*, and *bootstrap*. Details of these methods are found in [5].

In many studies, numerical comparisons of error rate estimators have been performed using bias, standard deviation and mean squared error as measures of goodness of the error rate estimators (e.g., [4], [8], [10]). Moreover, populations have been assumed to be normally distributed.

However, in analyses of real data sets, the assumption of a normal distribution may not be appropriate. Therefore, it is important to examine the behavior of error rate estimators under non-normal distribution. The commonly used measures of deviation from normality are skewness and kurtosis; however, their effects for the estimators have not been studied sufficiently.

Hence we focus on effects of these two measures of non-normality for error rate estimators. We concentrate on only kurtosis, which is generally easier to treat than skewness. A robustness study of the estimators is conducted by plotting

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 340–347, 2009.

the estimates obtained by the Monte Carlo method and by the Kruskal-Wallis test for means of biases and standard deviations of estimators corresponding to several kurtosises. Then we compare our results with results of preceding studies.

2 Definitions of Error Rates and Estimators

2.1 Error Rates

First we define the error rate in two-class discrimination. Let a set of cases $x_1, \ldots, x_n$ be available. This set is called a *training set* and is denoted by χ. Each case of χ is composed of the p-dimensional feature vector $\boldsymbol{t}_i$ and the class label $y_i \in \{\omega_1, \omega_2\}$. Moreover, let $r(\cdot|\chi)$ denote a discrimination rule based on χ, and $r(\boldsymbol{t}|\chi)$ gives the specific result of discrimination for an observation whose feature vector and unknown class label are $\boldsymbol{t}$ and y, respectively. The most general loss function of a discrimination rule is

$$Q[y, r(\boldsymbol{t}|\chi)] = \begin{cases} 0 & \text{if } \; y = r(\boldsymbol{t}|\chi), \\ 1 & \text{if } \; y \neq r(\boldsymbol{t}|\chi). \end{cases} \tag{1}$$

Let $\boldsymbol{T}$ and Y be random variables corresponding to $\boldsymbol{t}$ and y. If the joint distribution of $\boldsymbol{T}$ and Y is F, *the actual error rate* for $r(\cdot|\chi)$, denoted by e_{A}, is defined as

$$e_{\mathrm{A}} = E_F \left[Q[Y, r(\boldsymbol{T}|\chi)] \right], , \tag{2}$$

where $E_F[\cdot]$ represents an expectation with respect to F.

In addition we now define *the Bayes error rate*, denoted by e_{B}. Its mathematical definition is

$$e_{\mathrm{B}} = \int \int \left[1 - \max_i p(\omega_i|\boldsymbol{t}) \right] f(\boldsymbol{t}, y) d\boldsymbol{t} dy, \tag{3}$$

where $f(\cdot, \cdot)$ is the probability density function of F and $p(\omega_i|\boldsymbol{t})$ is the posterior probability of class ω_i given $\boldsymbol{t}$. The Bayes error rate is the minimum possible error rate given a set of features, and therefore $e_{\mathrm{A}} \geq e_{\mathrm{B}}$.

The objective of this study is a comparison of estimators of e_{A} with consideration given to effects of kurtosis. In order to compare estimators, it is necessary to determine the discrimination rule. We use Fisher's linear discriminant function (LDF), which assumes homoscedasticity, because it is a basic and the most commonly applied rule. When χ consists of samples drawn from ω_i and each sample size is n_i, LDF is defined as

$$l(\boldsymbol{t}|\chi) = (\bar{\boldsymbol{t}}_1 - \bar{\boldsymbol{t}}_2)' S^{-1} \boldsymbol{t} - \frac{1}{2} (\bar{\boldsymbol{t}}_1 - \bar{\boldsymbol{t}}_2)' S^{-1} (\bar{\boldsymbol{t}}_1 + \bar{\boldsymbol{t}}_2), \tag{4}$$

where

$$\bar{\boldsymbol{t}}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \boldsymbol{t}_{ij}, \quad S = \frac{1}{n_1 + n_2 - 2} \sum_{i=1}^{2} \sum_{j=1}^{n_i} (\boldsymbol{t}_{ij} - \bar{\boldsymbol{t}}_i)(\boldsymbol{t}_{ij} - \bar{\boldsymbol{t}}_i)'.$$

An observation x whose feature vector is $\boldsymbol{t}$ is classified to ω_1 if $l(\boldsymbol{t}|\chi) \geq 0$ and to ω_2 otherwise. See [1] for details of LDF.

2.2 Apparent Error Rate and Leave-One-Out Method

Among the estimators of e_{A}, we are interested in resampling-based estimators. Estimators based on bootstrap introduced by Efron [3] have been widely studied (e.g., [2], [10]). In this section we describe resampling-based estimators and bootstrap-based estimators using a training set χ of size n.

The apparent error rate, denoted by $\hat{e}_{\mathrm{app}}$, is

$$\hat{e}_{\mathrm{app}} = \frac{1}{n}\sum_{i=1}^{n} Q[y_i, r(\boldsymbol{t}_i|\chi)]. \tag{5}$$

This is the proportion of misclassified cases when we apply a discrimination rule $r(\cdot|\chi)$ to its own training set.

The apparent error rate usually underestimates e_{A} because the discrimination rule is constructed so that it performs well on the training set. Therefore, it is expected that this problem can be avoided by separating an available set into two sets: a set for training and a set for evaluating, namely, a *test set. The leave-one-out estimator*, denoted by $\hat{e}_1$, is an estimator based on this idea:

$$\hat{e}_1 = \frac{1}{n}\sum_{i=1}^{n} Q[y_i, r(\boldsymbol{t}_i|\chi^{(-i)})], \tag{6}$$

where $\chi^{(-i)}$ denotes the training set in which the case $(\boldsymbol{t}_i, y_i)$ is removed.

2.3 Normal Bootstrap Estimators

Bootstrap is generally used for estimating the bias of $\hat{e}_{\mathrm{app}}$ numerically under the empirical distribution, $\hat{F}$, which assigns mass $1/n$ to each case of χ. Let $\chi^* = \{(\boldsymbol{t}_i^*, y_i^*);\ i = 1, \ldots, n\}$ be a random sample of size n from $\hat{F}$. The actual error rate and apparent error rate for a discrimination rule based on χ^* are

$$e_{\mathrm{A}}^* = \frac{1}{n}\sum_{i=1}^{n} Q[y_i, r(\boldsymbol{t}_i|\chi^*)], \quad \hat{e}_{\mathrm{app}}^* = \frac{1}{n}\sum_{i=1}^{n} Q[y_i^*, r(\boldsymbol{t}_i^*|\chi^*)].$$

The bias of $\hat{e}_{\mathrm{app}}^*$ is the expectation of $(e_{\mathrm{A}}^* - \hat{e}_{\mathrm{app}}^*)$ and its approximation can be computed by the Monte Carlo method as follows. We generate $\chi^*(1), \ldots, \chi^*(B)$, which are called bootstrap samples, with replacement from $\hat{F}$ and compute $(e_{\mathrm{A}}^*(b) - \hat{e}_{\mathrm{app}}^*(b))$ based on the bth bootstrap sample $(b = 1, \ldots, B)$. The average of those values is the approximation of the bias:

$$bias^{(\mathrm{bs})} = \frac{1}{B}\sum_{b=1}^{B} \left\{ e_{\mathrm{A}}^*(b) - \hat{e}_{\mathrm{app}}^*(b) \right\}. \tag{7}$$

The bias of $\hat{e}_{\text{app}}$ is estimated by $bias^{(\text{bs})}$. In this way, *the bootstrap estimator*, denoted by $\hat{e}_{\text{bs}}$, is introduced:

$$\hat{e}_{\text{bs}} = \hat{e}_{\text{app}} + bias^{(\text{bs})}. \tag{8}$$

The actual error rate is interpreted as the error rate for observations whose distance from the training set χ is zero. However, future observations may have various distances from χ. Therefore, an adoptively weighted linear combination of $\hat{e}_{\text{app}}$ and an estimator of the error rate for observations away from χ is expected to be an unbiased estimator. The latter estimator based on bootstrap is

$$\hat{e}_{\text{b0}} = \frac{1}{n}\sum_{b=1}^{B}\sum_{i=1}^{n}\left\{\frac{1}{B_i}\sum_{b\in C_i} Q[y_i, r(\boldsymbol{t}_i|\chi^*(b))]\right\}, \tag{9}$$

where C_i is the set of numbers of the bootstrap samples that do not include case $(\boldsymbol{t}_i, y_i)$ and B_i is the size of C_i. Efron [3] showed that the following estimator, which is called .632 estimator, is almost unbiased under some assumptions:

$$\hat{e}_{.632} = 0.368\,\hat{e}_{\text{app}} + 0.632\,\hat{e}_{\text{b0}}. \tag{10}$$

2.4 Smoothed Bootstrap Estimators

If F is known to be continuous, we should use a smoothed distribution function instead of $\hat{F}$ in the bootstrap procedure. The randomized bootstrap estimator, which was also proposed by Efron [3], is an estimator that is smoothed in the class membership direction. We draw bootstrap samples from the distribution $\hat{F}_{\text{s}}$, which assigns mass p_i/n to $(\boldsymbol{t}_i, y_i)$ and mass $(1-p_i)/n$ to $(\boldsymbol{t}_i, \bar{y}_i)$. Now $\bar{y}_i$ is the class label opposite to y_i. Then we estimate the bias of $\hat{e}_{\text{app}}$ in the same way as (7).

Efron [3] pointed out that $l(\boldsymbol{t}_i|\chi)$ was naturally associated with

$$p_i = \frac{1}{1+\exp\{(-1)^{y_i} l(\boldsymbol{t}_i|\chi)\}} \qquad (i = 1, \ldots, n). \tag{11}$$

We call the estimator using this p_i *the randomized bootstrap estimator*, and it is denoted by $\hat{e}_{\text{rb}}$. On the other hand, the estimator using $p_i = 0.9$ is called *the simple randomized bootstrap estimator* and is denoted by $\hat{e}_{\text{srb}}$.

Another smoothed bootstrap estimator is *the convex bootstrap estimator* (Chernick et al. [2]), $\hat{e}_{\text{cb}}$. Sampling two cases $(\boldsymbol{t}_i, y_i)$ and $(\boldsymbol{t}_j, y_j)$ from χ with $y_i = y_j$, we generate a new case

$$(\boldsymbol{t}, y) \quad [\boldsymbol{t} = \lambda\boldsymbol{t}_i + (1-\lambda)\boldsymbol{t}_j,\ y = y_i = y_j], \tag{12}$$

where λ is a uniform random number on $[0, 1]$. For each class, this process is repeated the same number of times as the number of cases included in χ. The generated cases are used as bootstrap samples, and latter process of calculating approximation of the bias of $\hat{e}_{\text{app}}$ is the same as (7).

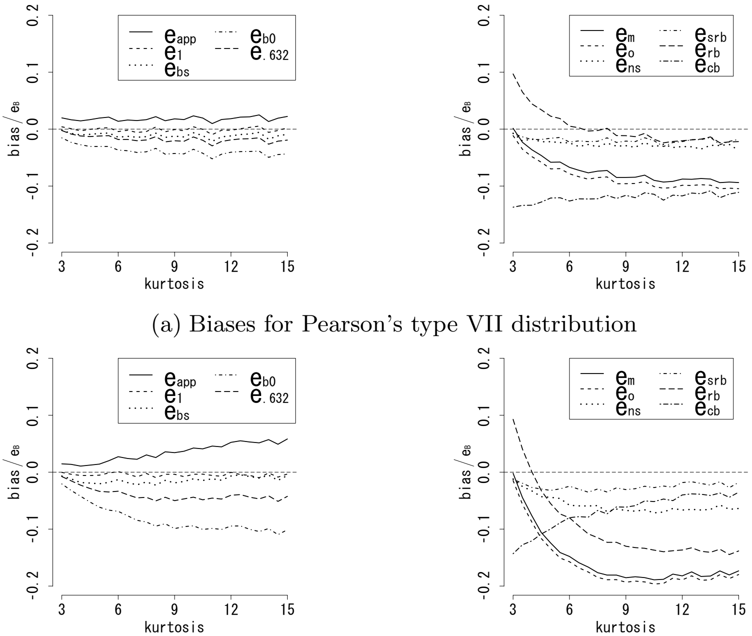
3 Simulation Study

For estimating the actual error rates of classification rules, it is desirable that the estimator is unbiased and its standard deviation is as small as possible. In addition, the bias and standard deviation of the estimator should be robust to kurtosis. Therefore, in this section we calculate the bias and standard deviation of estimators by the Monte Carlo method and discuss the performance and robustness of them by plotting the calculated values. Furthermore, in order to discuss the robustness quantitatively, we carry out the Kruskal-Wallis test for means of biases and standard deviations of estimators corresponding to several kurtosises.

3.1 Conditions of Simulation

We first compare the performances of eight estimators described in Section 2. Next, the behaviors of McLachlan's estimator [6], Okamoto's estimator [7], and the NS estimator [9], which are denoted by $\hat{e}_{\mathrm{m}}$, $\hat{e}_{\mathrm{o}}$, and $\hat{e}_{\mathrm{ns}}$, respectively, are also examined. The estimators $\hat{e}_{\mathrm{m}}$ and $\hat{e}_{\mathrm{o}}$ are derived under the assumption of normality, and $\hat{e}_{\mathrm{ns}}$ is $\hat{e}_{\mathrm{app}}$ with a smoothed loss function using the distribution function of a standard normal distribution.

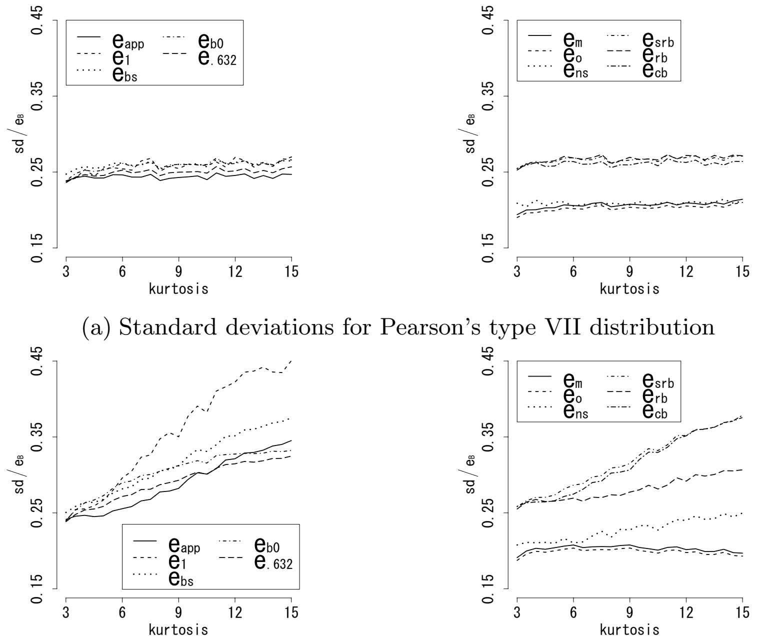
In the calculation of bias and standard deviation, we assume that the shapes of population distributions of classes are identical but that the means are different.



(a) Biases for Pearson's type VII distribution

(b) Biases for exponential power distribution

Fig. 1. Comparison of biases ($e_{\mathrm{B}} = 0.3$)



(a) Standard deviations for Pearson's type VII distribution

(b) Standard deviations for exponential power distribution

Fig. 2. Comparison of standard deviations ($e_{\mathrm{B}} = 0.3$)

The conditions for the calculation are as follows: the sample size of each class is 20, the common standard deviation is one, the prior probability of each class is identical, the number of simulation trials is 500, and the number of bootstrap samples is 100.

Increasing the kurtosis from 3 to 15 with steps of 0.5, we calculate the bias and standard deviation in the case that the Bayes error rate is e_{B} under the above conditions. We therefore need a distribution for which we can determine its kurtosis regardless of the values of other parameters. Pearson's type VII distribution and exponential power distribution are examples of such distributions and we use them as the population distributions.

3.2 Simulation Results

We first discuss the results for bias. Figure 1 shows two plots of ratios of the biases of estimators to e_{B} for populations (a) Pearson's type VII distribution and (b) exponential power distribution when e_{B} is 0.3. To save space, other cases are omitted. In (a) and (b), the estimators plotted in the left graph are more robust to kurtosis and have smaller bias than those in the right graph. $\hat{e}_1$ and $\hat{e}_{\mathrm{bs}}$ are almost unbiased against the change in kurtosis, and therefore these two estimators may be superior to $\hat{e}_{.632}$ in terms of bias, though $\hat{e}_{.632}$ has been highly evaluated in preceding studies. Another remarkable point is the large difference between the results for (a) and (b). This indicates that the kurtosis

Table 1. Robustness of estimators for kurtosis

(a) Based on bias

	e_{B}			
Est	0.1	0.2	0.3	0.4
$\hat{e}_{\mathrm{app}}$	×	○	○	×
$\hat{e}_1$	○	○	○	○
$\hat{e}_{\mathrm{bs}}$	○	○	×	×
$\hat{e}_{\mathrm{b0}}$	×	×	×	○
$\hat{e}_{.632}$	×	○	×	×
$\hat{e}_{\mathrm{m}}$	×	×	×	×
$\hat{e}_{\mathrm{o}}$	×	×	×	×
$\hat{e}_{\mathrm{ns}}$	×	×	×	×
$\hat{e}_{\mathrm{srb}}$	○	○	×	×
@ $\hat{e}_{\mathrm{rb}}$	×	×	×	×
$\hat{e}_{\mathrm{cb}}$	×	×	×	×

(b) Based on standard deviation

	e_{B}			
Est	0.1	0.2	0.3	0.4
$\hat{e}_{\mathrm{app}}$	○	○	○	×
$\hat{e}_1$	○	○	×	×
$\hat{e}_{\mathrm{bs}}$	○	×	×	×
$\hat{e}_{\mathrm{b0}}$	×	×	×	○
$\hat{e}_{.632}$	×	×	×	○
$\hat{e}_{\mathrm{m}}$	×	×	×	×
$\hat{e}_{\mathrm{o}}$	×	×	×	×
$\hat{e}_{\mathrm{ns}}$	×	×	○	○
$\hat{e}_{\mathrm{srb}}$	○	×	×	×
@ $\hat{e}_{\mathrm{rb}}$	×	×	○	×
$\hat{e}_{\mathrm{cb}}$	×	×	×	×

of the distribution will affect the bias, but the amount of change is different depending on the population distribution.

Next, we examine the results for standard deviation. Figure 2 shows plots of the ratios of standard deviations of the estimators to e_{B} for (a) and (b) when e_{B} is 0.3. As in the case of bias, values in the same kurtosis are different depending on the distribution. Moreover, in most cases, resampling-based estimators have larger standard deviation than do $\hat{e}_{\mathrm{m}}$, $\hat{e}_{\mathrm{o}}$, and $\hat{e}_{\mathrm{ns}}$. Although smoothed bootstrap estimators are used for the purpose of making their standard deviations be smaller than those of conventional resampling-based estimators, smoothed bootstrap estimators have the same or larger standard deviation than those of other resampling-based estimators.

For a more quantitative discussion of robustness of the estimators, we carry out the Kruskal-Wallis test. We examine whether there is a significant difference among the means of biases and standard deviations of an estimator for several kurtosises by using the test. The test is done at significance level $\alpha = 0.05$ and we decide that the estimator is robust to kurtosis if a significant difference does not exist. Table 1 shows the results for Pearson's type VII distribution when sample size of each class is 20, and the results are represented by $\circ$ if the estimator is robust in that case and by $\times$ otherwise. The test is performed for kurtosis $k = 3, 6, 9, 12, 15$ using 40 biases and standard deviations calculated in each k.

Our results show that $\hat{e}_1$ is always robust to kurtosis based on bias and that normal bootstrap estimators are more robust than smoothed bootstrap estimators on the whole. Based on standard deviation, it is difficult to find an estimator judged to be robust. Moreover, from our numerical results, the estimators might be classified into two groups. $\hat{e}_{\mathrm{app}}$, $\hat{e}_1$, $\hat{e}_{\mathrm{bs}}$, and $\hat{e}_{\mathrm{srb}}$ are robust when e_{B} is small, and the other estimators are robust when e_{B} is large.

4 Concluding Remarks

We compared several estimators for e_{A} in two-class discrimination, especially resampling-based estimators, from the viewpoint of robustness to kurtosis. Our results show that $\hat{e}_1$ is almost unbiased regardless of the value of kurtosis. In addition, $\hat{e}_{.632}$, which has been reported to be superior to other estimators, is sensitive to kurtosis and is inferior to $\hat{e}_1$ and $\hat{e}_{\mathrm{bs}}$ from the aspect of bias. On the other hand, smoothed bootstrap estimators have larger standard deviation than do other resampling-based estimators.

Our results suggest the necessity of considering non-normality in comparison of error rate estimators. However, we considerd the effects of kurtosis only when using LDF in a one-dimensional distribution in this paper. Therefore, It is necessary to examine the effects of various non-normalities when using other discrimination rules in multidimensional distributions.

References

[1] Anderson, T.W.: Classification by Multivariate Analysis. Psycometrika 16, 631–650 (1951)

[2] Chernick, M.R., Murthy, V.K., Nealy, C.D.: Application of Bootstrap and Other Resampling Techniques: Evaluation of Classifier Performance. Pattern Recognition Letters 3, 167–178 (1985); 4, 133–142 (1986)

[3] Efron, B.: Estimating the Error Rate of a Prediction Rule: Improvement on Cross-validation. Journal of the American Statistical Association 78, 316–331 (1983)

[4] Ganeshanandam, S., Krzanowski, W.J.: Error-rate Estimation in Two-group Discriminant Analysis Using the Linear Discriminant Function. Journal of Statistical Computation and Simulation 36, 157–175 (1990)

[5] Hand, D.J.: Recent Advances in Error Rate Estimation. Pattern Recognition Letters 4, 335–346 (1986)

[6] McLachlan, G.J.: An Asymptotic Unbiased Technique for Estimating the Error Rates in Discriminant Analysis. Biometrics 30, 239–249 (1974)

[7] Okamoto, M.: An Asymptotic Expansion for the Distribution of the Linear Discriminant Function. Annals of Mathematical Statistics 34, 1286–1301 (1963)

[8] Page, J.T.: Error-Rate Estimation in Discriminant Analysis. Technometrics 27, 189–198 (1985)

[9] Snapinn, S.M., Knoke, J.D.: An Evaluation of Smoothed Classification Error-rate Estimators. Technometrics 27, 199–206 (1985)

[10] Wehberg, S., Schumacher, M.: A Comparison of Nonparametric Error Rate Estimation Methods in Classification Problems. Biometrical Journal 46, 35–47 (2004)

Reasoning about External Environment from Web Sources

Hércules Antonio do Prado[1,2], André Ribeiro Magalhães[1], and Edilson Ferneda[1]

[1] Graduate Program on Knowledge and IT Management, Catholic University of Brasilia, Brasília, DF, Brazil
[2] Embrapa – Management and Strategy Secretariat, Brasília, DF, Brazil
hercules@{ucb.br,embrapa.br}, ribeiromagalhaes@gmail.com, eferneda@pos.ucb.br

Abstract. Most organizations approach internal and external challenges with a varied degree of effectiveness. One of their biggest challenges is the ability to identify and respond appropriately to changes in their external environments. These changes affect not only their technological choices, but also their internal structures and cultures. In this context, we have seen an increasing demand for computational tools capable not only to support information storage but also to help in reasoning about the organizational environment. In particular, it is observed that the availability of a huge set of information in the Web offers a new opportunity to learn and reason about the organizational context. In this paper we present an empirical model to proceed the knowledge extraction from Web sources and support the reasoning process in the Competitive Intelligence domain.

Keywords: Informal Reasoning, Competitive Intelligence, Environmental Scanning.

1 Introduction

The increasing availability of information on the Web is leading to several changes in our society (GANTZ, 2008; NEGROPONTE, 1995). In businesses, the Web has been used to reach current and potential customers. But it has also served as a source of information for professionals involved in environmental scanning. Intelligence analysts have been seeking for information on the Web to support decision making in organizations.

With the increasing development of Information and Communication Technologies, there is a rapid obsolescence of mechanisms for collecting information on the Web (CHOO et al., 2000; TEO and CHOO, 2001). Thus, the intelligence analysis demands new sophisticated tools for monitoring the Web in the search for information that can support the analyst in reasoning about the external environment.

This paper presents W3EnvScan (World Wide Web Environment Scanning), a tool for monitoring the Web able to identify information changes in heterogeneous representation, like HTML, Ajax, and XML pages, or developed in ASP, Java, Perl, PHP, Python, or Ruby. Some practical examples of how to use the tool are also presented.

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 348–355, 2009.

2 Reported Works

Monitoring changes in a competitive economy characterized by globalization is a critical factor for success in organizations. These changes impose strong influences of the external environment in its decision making processes. With the increasing availability of information on the web, monitoring the external environment via the Internet has become a relevant issue to the intelligent organization.

Flexibility, agility, and simple market operational structures are essential characteristics for success and survival of organizations. The identification of changes in the external environment in order to plan the required interventions in the internal environment is known as Competitive Intelligence (CI).

As an important source of strategic information, the Web promotes technological developments and, consequently, the creation of sophisticated techniques of representation and storage of information. This context demands new tools for information search and retrieval.

The Web has never stopped evolving and the emergence of Web 2.0 certainly helped to consolidate a perception of value from what new business models can benefit from. Also, this scenery led to the emergence of new sources of revenue generation and cost reductions by means of tools for environmental scanning like W3EnvScan.

Guimarães (2006) emphasizes that "the external environment is a source of resources from where organizations extract the information needed to interact, adopting positive changes, adapting to the negative influences, or alleviating their effects." He conducted a qualitative study on twelve companies of Informatics and Telecommunications involved with environment scanning (ES) in the CI process. He also applied a set of techniques and tools to identify the market trends. The author notices that: *(i)* Internet arises as a primary source of information to the organization, spreading the use of ES tools, and *(ii)* even though many companies do not follow the guidelines of CI, at some point they implement some of the stages of this process.

Thomé (2006) presents a study to identify the requirements for CI tools in an agricultural research company. The *Autonomy*[1] environment is presented as an alternative to automate the synthesis and the treatment of unstructured information. Another environment for CI, the *Córtex Intelligence*[2], is discussed. It uses robots to monitor for automated search, allowing the automation of a series of activities of CI by using Text Mining. The work emphasizes CI as a facilitator for decision making, improving the quality of any branch of knowledge.

Based on the results presented by Silva (2000), Lemos (2005) presents a model of multiagents to support the process of CI. The author proposes a multiagent architecture for a tool capable of monitoring the content of Web pages by applying techniques of textual analysis and detecting changes and the relevance of these changes.

[1] http://www.autonomy.com/

[2] http://www.cortex-intelligence.com/html/solucoes/plataforma.html

3 The W3EnvScan Model

One of the objectives of the monitoring model proposed is to provide an easy alternative to monitor a set of variables from the external environment. This independence enables the user to assess the real needs when choosing the parameters to be used. The model is presented by means of a functional representation (Fig. 1).

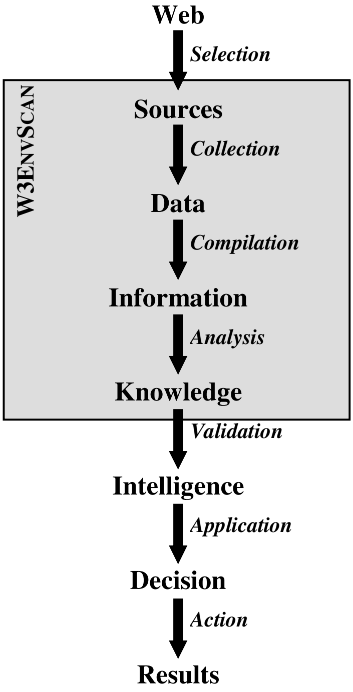


Fig. 1. W3EnvScan and its context

This representation includes the following steps:

- In the Knowledge phase the information goes through a validation process, with focus on the user objectives. The information obtained is compared with the target and its importance against the strategic benchmark is evaluated.
- In the Intelligence phase, managers apply their ability, business skills, and experience in the organization to identify strategic alternatives, such as: *(i)* new research projects, *(ii)* cooperation agreements, *(iii)* technology transference, and *(iv)* actions and reactions from the competition.
- In the Decision phase, the intelligence analysts, along with managers, assess the results for strategies, seeking to extract the information for decision making.

- In the last phase, the results are evaluated. This phase will provide indicators on all the previous steps and, if well used, will improve the decision-making process in the organization.

The W3EnvScan is based on this model and includes the characteristics of portability, beyond a high degree of scalability. Figure 2 represents the components of W3EnvScan. The system makes the monitoring of Web pages previously registered. This information is stored in a persistent repository that can be accessed for information retrieval by the intelligence analyst.

W3EnvScan is able to perform the real time access to pages registered, reporting failures in case of unavailability. When accessing a page, some checking is carried out and information regarding the page is generated; this information can be used for analysis and decision making. As changes occur in the monitored pages, intelligence analysts will be promptly informed so that they can take the appropriate measures according to the changes.

To carry out the verification and generation of pages of information tracked, it was implemented a multiagent architecture. Functions were implemented that correspond to the following components: the user interface, controller, and interface with the environment. The service functions *Maintain*, *Maintain RSS*, *Keep Track*, and browser service have characteristics of components of the *User Interface* type. A component controller was implemented with the function *Maintain Database* and a component interface with two functions, *RSS Keep* and *Track Service*. It should be noted that *RSS Keep* is mapped into two components (*Interface with the environment* and the *User interface*) as well as *Monitor service*. This is due to the fact that both functions acquire information about the external environment to present the results of intelligence analysts.

Although the availability of new methods for information processing that have made CI a powerful tool, the professional involved in the transformation process of knowledge still represents an important factor for the success of any environmental scanning. Ultimately, the analyst is the one that will reason about the Web. He seeks for increasing organizational competitiveness based on information collected and

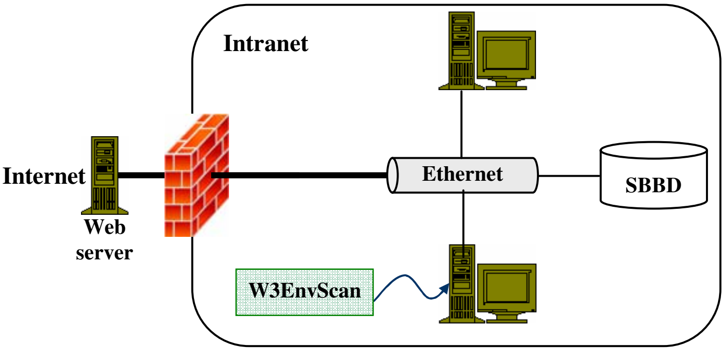


Fig. 2. W3EnvScan structure for collecting and monitoring

stored through the Competitive Intelligence System (CIS). He became an important source of information, often unexplored, due to the difficulties inherent in this large volume of information.

Beyond monitoring pages in Web, W3EnScan provides features that enable the analyst to acquire patterns from those data. These features include trend analysis, alarms, and classifiers that can be applied over the persisted data. The mining task is parameterized when the analyst defines a page or a set of pages to be monitored.

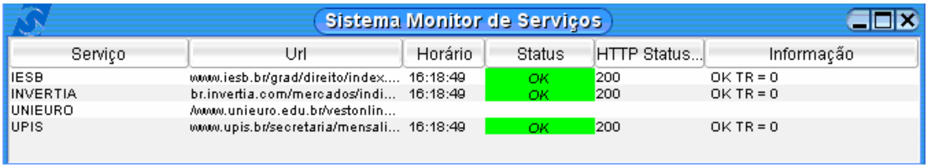
4 Examples of Application

This section presents how W3EnvScan monitors variables and some possible patterns that can be discovered by applying the features available in the tool of monitoring.

4.1 Monitoring Variables

The situation addressed in this case study is a college that is launching an undergraduate course on Laws. The managers are interested in monitoring the variation in the course prices of the competitors. By monitoring monthly the behavior of market with CI techniques, the college intends to maintain a balance in the price-quality relation, while keeping an adequate number of students enrolled.

We considered three competitors to be monitored: Institute of Higher Education of Brasília (Iesb)[3], Paulista University (Upis)[4] and Unieuro College[5]. The process begins with the intelligence analyst introducing the information related to the pages to be monitored. After the specification of that information, the system starts monitoring the pages, showing a panel in which the page status is depicted (Fig. 3).



Sistema Monitor de Serviços

Serviço	Url	Horário	Status	HTTP Status...	Informação
IESB	www.iesb.br/grad/direito/index....	16:18:49	OK	200	OK TR = 0
INVERTIA	br.invertia.com/mercados/indi...	16:18:49	OK	200	OK TR = 0
UNIEURO	/www.unieuro.edu.br/vestonlin...				
UPIS	www.upis.br/secretaria/mensali...	16:18:49	OK	200	OK TR = 0

Fig. 3. Monitoring with W3EnvScan

The monitoring is carried out by means of visual alerts that indicate the page status; it can be *normal access* or *failure*. The green color means that the page has been accessed, but no change was detected, while red color indicates that the page was modified. If a modification occurs in a marked space of the page, the status of the service is changed to critical and the line is set to red color. So, the analyst can access the modified page. This simple application can be used by an organization to set their prices according to its competitors' levels.

[3] Iesb College (http://www.iesb.br)

[4] Paulista University (http://www.upis.br)

[5] Unieuro College (http://www.unieuro.edu.br)

4.2 Generating Knowledge

W3EnvScan provides a set of features that enable the intelligence analyst to obtain patterns as rules, tendencies, and alarms. Next, an example of the mining approaches that can be applied by means of W3EnvScan is depicted.

In a general plot, one organization can be interested in explaining how some variables are related to make well-informed decisions. For example, one may need to know how a stock market index (dependent variable), say the IBOVESPA, in time *t+1*, is affected by other indexes (independent variables) around the world in time *t*. The analyst can specify as independent variables the indexes NASDAQ (USA), DOW JONES (USA), NIKKEI (Asia), DAX (German), and MERVAL (Argentina) to be monitored by W3EnvScan and, periodically, obtain the classification rules that show the relations among them and the dependent variable. Table 1 shows the ups and downs in the indexes in a time-frame of 6 days.

Table 1. Variations of 6 stock market indexes for 6 days[6]

Date	**Nasdaq**	**Dow Jones**	**Nikkei**	**DAX**	**MERVAL**	IBOVESPA
March 05	Down	Down	Down	Down	Down	Down
March 06	Down	Up	Down	Down	Up	Up
March 09	Down	Down	Down	Up	Down	Up
March 10	Up	Up	Up	Up	Up	Up
March 11	Up	Up	Down	Up	Steady	Up
March 12	Up	Up	Up	Up	Up	Down

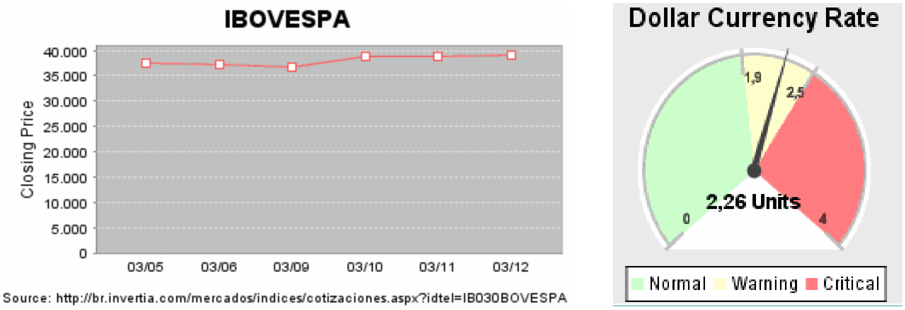
If the analyst chooses the rule generator CNM (Combinatorial Neural Model) (MACHADO, 1990), the system would show him, for example, the rule:

If DOW JONES **is** UP **then** IBOVESPA **is** UP (Confidence = 75%, Support = 50%)

Confidence and support are metrics that enable the quality evaluation of the rule. Confidence expresses the amount of rules satisfying the right side among the rules that satisfy the left side. Support is the percentage of examples for what the rule is held.

Other features include are *Trend Analysis* (Fig. 4a) that enables the analyst to assess how a variable changes along the time and *Alarm* (Fig. 4b) that inform the analyst when a variable monitored reaches a specified value. Figure 4a refers to a monitoring in which the intelligence analyst specified as objects to be monitored the daily closing prices of a stock in a stock and Figure 4b shows an alarm based on the variation of the dollar rate. The red range in the control represents the point in which the dollar reaches the specified limit. For historical series, the analyst may set an option to build a predictive model based on the evolution of the stock market over time. When triggering the analysis, the parameters related to the chosen technique shall be required from the analyst in order to execute the requested time series analysis.

[6] http://br.invertia.com/mercados/indices/



Source: http://br.invertia.com/mercados/indices/cotizaciones.aspx?idtel=IB030BOVESPA

(a) *(b)*

Fig. 4. Examples of *(a)* time series and *(b)* alarm

5 Conclusion and Further Works

The use of the RSS service proved to be partially effective. If, on the one hand, the RSS informed the occurrence of changes in a page, on the other it does not specify what changed in that page. To overcome this limitation, the HTML code from a recovered page must be processed to identify the location of the change and capture the new value. For this purpose, different versions of a page are kept.

Due to a lack of standardization in the development of HTML pages, the processing of HTML code needs to be substantially improved to handle more complex display objects, such as fields, text or animated frames.

In the evolution of this work it is going to be included control panels that represent the action to be carried out under certain conditions.

During the tests three kinds of information should be considered while monitoring the Web: *(i)* the mere information of content change in a page, *(ii)* the details about changes occurred in a page, and *(iii)* the result of related information that has changed in more than one page. W3EnvScan model covers completely the first kind and, partially, the second one. To cover the third kind of information, it requires more improvements.

In general, we observed that the monitoring of the Web is an interesting alternative for the detection of changes in the external environment of organizations, requiring, however, significant improvements. The Semantic Web is an example of technological change that opens up prospects for a better use of this framework of information to allow standardization in the construction of pages and more focused information monitoring.

References

Choo, C.W., Detlor, B., Turnbull, D.: Web work: Information seeking and knowledge work on the World Wide Web. In: Information Science & Knowledge Management. Springer, Heidelberg (2000)

Gantz, J.F., et al.: The diverse and exploring digital universe. IDC White Paper (2008), http://www.emc.com/collateral/analyst-reports/diverse-exploding-digital-universe.pdf

Guimarães, C.: Study on the use of external information to decision support – a general view on Informatics companies of Belo Horizonte. MSc Dissertation in Information Science, Federal University of Minas Gerais, Belo Horizonte, Brazil (2006), http://www.bibliotecadigital.ufmg.br/ (in Portuguese)

Lemos, A.F.: Scanning the information source in Internet: multiagent model to support the Competitive Intelligence process. MSc Dissertation in Production Engineering, Federal University of Santa Catarina, Florianópolis (Brazil) (2005), http://teses.eps.ufsc.br/ (in Portuguese)

Machado, R.J., da Rocha, A.F.: The Combinatorial Neural Network: A Connectionist Model for Knowledge Based Systems. In: Bouchon-Meunier, B., Zadeh, L.A., Yager, R.R. (eds.) IPMU 1990. LNCS, vol. 521, pp. 578–587. Springer, Heidelberg (1991)

Negroponte, N.: Being Digital. Alfred Knopf, Inc., New York (1995)

Silva, H.P.: Competitive intelligence in Internet: a process proposal. PhD Thesis in Production Engineering, Federal University of Santa Catarina, Florianópolis (Brazil) (2000), http://teses.eps.ufsc.br/defesa/pdf/1750.pdf (In Portuguese)

Teo, T.S.H., Choo, W.Y.: Assessing the impact of using the Internet for competitive intelligence. Information & Management 39, 67–83 (2001)

Thomé, M.F.: A tool for competitive intelligence support: a case study in Embrapa. MSc Dissertation in Knowledge and Information Technology Management, Catholic University of Brasília, Brasília, Brazil (2006), http://www.bdtd.ucb.br/tede/ (In Portuguese)

An Agent Control Method Based on Variable Neighborhoods

Seiki Ubukata[1], Yasuo Kudo[2], and Tetsuya Murai[1]

[1] Hokkaido University, Kita 14, Nishi 9, Kita-ku, Sapporo 060-0814, Japan
{ubukata,murahiko}@main.ist.hokudai.ac.jp
[2] Muroran Institute of Technology, 27-1 Mizumoto-cho, Muroran 050-8585, Japan
kudo@csse.muroran-it.ac.jp

Abstract. In this paper, we propose a model that an agent selects actions based on variable neighborhoods. We formulate relationships among variable neighborhoods, the agent's observations, and the agent's behaviors in a framework of rough set theory and topological spaces. The main task is to explore a method by which we can select sizes of neighborhoods under given contexts. We also show simulation results of the proposed method.

Keywords: Rough set theory, Topological spaces, Modal logic, Kripke models, Neighborhood models, Variable neighborhoods, Agent control.

1 Introduction

In [3], we formulated a rough-set-based method of behavior arbitration in agent control using behavior-based AI principle and applied it to the agent's garbage collection. There we found some deadlock problems which could be solved by reducing agent's ranges of view. This observation suggests that we need further detailed topological consideration on the previous method. From a topological point of view, such solution can be interpreted as agent's control of sizes of his range of view as a neighborhood.

In this paper, we propose a variable neighborhood model that agents select actions based on variable neighborhoods. The agent can select one of neighborhoods in a neighborhood system under a given context. We examine how the way of selecting sizes of neighborhoods effects the agents' behaviors. We also show some simulation results of the proposed method applied to agent control in target chasing problems.

2 Preliminaries

2.1 Rough Set Theory

Given a non-empty set U, called a universe, and an equivalence relation R on U, a Pawlak approximation space is the pair (U, R). For a subset $X \subseteq U$, the lower and upper approximations of X are defined by, respectively,

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 356–363, 2009.

$$[R]X = \{x \in U \mid [x]_R \subseteq X\},$$
$$\langle R\rangle X = \{x \in U \mid [x]_R \cap X \neq \emptyset\},$$

where $[x]_R$ is the equivalence class of x with respect to R. The pair $([R]X, \langle R\rangle X)$ is called the rough set of X with respect to R. If X satisfies $[R]X = X = \langle R\rangle X$, X is said to be R-definable, else X is R-rough.

U is divided into the following three disjoint regions by the approximations:

- R-positive region of X: $POS_R(X) = [R]X$
- R-negative region of X: $NEG_R(X) = U \setminus \langle R\rangle X$
- R-boundary region of X: $BD_R(X) = \langle R\rangle X \setminus [R]X$.

2.2 Possible World Semantics for Modal Logics

There are a variety of possible world semantics in modal logic literature and, among them, Kripke-style semantics is well-known. A structure

$$\mathscr{F} = \langle U, R\rangle$$

is called a Kripke frame, where U is a universe (set of possible worlds) and R is a binary relation on U (accessibility relation). A Kripke frame with a valuation V, that is,

$$\mathscr{M} = \langle U, R, V\rangle$$

is called Kripke model. A valuation assigns either truth value 1 (true) or 0 (false) to every atomic sentence at each world.

An atomic sentence p is said to be true at a possible world $x \in U$ in a Kripke model $\mathscr{M}$, denoted $\mathscr{M}, x \models p$, if and only if $V(p, x) = 1$. Then, the truth conditions for modal sentences are given by

$$\mathscr{M}, x \models \Box p \Leftrightarrow U_R(x) \subseteq ||p||^{\mathscr{M}},$$
$$\mathscr{M}, x \models \Diamond p \Leftrightarrow U_R(x) \cap ||p||^{\mathscr{M}} \neq \emptyset,$$

where $||p||^{\mathscr{M}}$ and $U_R(x)$ are subsets in U defined by

$$||p||^{\mathscr{M}} = \{x \in U \mid \mathscr{M}, x \models p\},$$
$$U_R(x) = \{y \in U \mid xRy\},$$

respectively.

2.3 Topology

A topological space is defined as $\mathscr{F} = \langle U, N\rangle$, where U is a universe and $N : U \to 2^{2^U}$ is a neighborhood system which satisfies the following conditions: for all $x \in U$

(N$_1$) $U \in N(x)$
(N$_2$) $X \in N(x) \Rightarrow x \in X$
(N$_3$) $X_1, X_2 \in N(x) \Rightarrow X_1 \cap X_2 \in N(x)$
(N$_4$) $(X \in N(x)$ and $X \subseteq Y) \Rightarrow Y \in N(x)$
(N$_5$) $X \in N(x) \Rightarrow \exists Y \in N(x)[Y \subseteq X$ and $\forall y(y \in Y \Rightarrow X \in N(y))]$

Topological spaces can be also defined using, for example, a family of open sets, a family of closed sets, and so on.

A neighborhood model is a structure $\mathscr{M} = \langle U, N, V \rangle$, where V is a valuation. In this case, N dose not necessarily have to satisfy the above conditions of neighborhood systems.

3 Framework of the Proposed Method

3.1 Target Chasing Problems

In this paper, we deal with target chasing problems where we consider several kinds of objects such as an agent, targets, walls and a base in a universe U, which may be topological or distance spaces.

- An agent has mobility capability.
- An agent has energy level and loses energy by performing actions.
- The purpose of an agent is to chase targets.
- If an agent's energy level becomes 0, it will stop.
- In a base, an agent can charge energy.

We can classify the object in two groups. One group contains an agent, targets (target1, target2, $\cdots$) and walls. They are called Type A objects in this paper. The other one contains a base called Type B object in this paper. Two different Type A objects cannot occupy one location at the same time. Type A object and Type B object, or different Type B objects can exist at the same location at the same time. An agent and targets are located at a point x in U. To chase a target, an agent should avoid walls and charge energy if its energy is low.

In our approach, an agent can observe objects only in the neighborhood that the agent selects. When an agent is located at x in U his neighborhood is denoted by $U(x)$. An agent cannot necessarily observe all range of $U(x)$ by, for example, walls illustrated in Fig. 1.

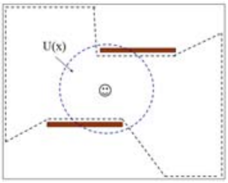


Fig. 1. An agent's range of view and walls

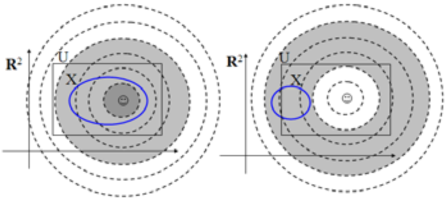


Fig. 2. The rough set of X by the fundamental system of neighborhood $B_{\alpha/n}$

3.2 The 2-Dimensional Euclidean Space and Rough Sets

The 2-dimensional Euclidean space $(\mathbb{R}^2, d)$ is a topological space. We discuss a way of generating rough sets from neighborhood systems. We assume that an agent is in a topological subspace (U, N_U) of the 2-dimensional Euclidean space, where $U \subseteq \mathbb{R}^2$ is a bounded closed set, N_U is a neighborhood system induced from $(\mathbb{R}^2, d)$.

We consider an α/n-open neighborhood

$$B_{\alpha/n}(x) = \{y \in \mathbb{R}^2 \mid d(x, y) < \alpha/n, n \in \mathbb{N}\}.$$

Selecting different α generates a variety of granularity. We can obtain disjoint classes by considering difference sets of α/n-open neighborhoods, The difference set $C_n^\alpha = B_{\alpha/n}(x) \setminus B_{\alpha/(n+1)}(x)$ becomes an annulus. Then the following set

$$P^\alpha = \{C_n^\alpha \mid n \in \mathbb{N}\} \cup \{U \setminus B_\alpha, \{x\}\}$$

is a partition of U, therefore we can define the equivalence relation R_α on U.

Given a subset $X \subseteq U$, we can have the rough set $([R_\alpha]X, \langle R_\alpha \rangle X)$ of X. In cases that the upper approximation of X is not a subset of U, the rough set is modified $([R_\alpha]X, (\langle R_\alpha \rangle X) \cap U)$.

An agent located at x decides the value of n and its neighborhood. As agent's criteria for selecting neighborhoods, we consider adopting the lower approximation, adopting the upper approximation, and so on. An agent can have a neighborhood $B_{\alpha/n}(x)$ by selecting n so that it can satisfy a condition, like $B_{\alpha/n}(x) \subseteq [R_\alpha]X$ and $B_{\alpha/n}(x) \subseteq \langle R_\alpha \rangle X$, depending on a given context.

3.3 Dynamic Neighborhood Model

In this subsection, we introduce a kind of neighborhood models in order to describe interaction between an agent and dynamic environment. We call them dynamic neighborhood models. We confine ourselves only to one agent cases in this paper. Let $x^t \in U$ be an agent's location at time t and let $\alpha(t)$ be a real number which an agent determines under a given context at time t. A dynamic neighborhood model is defined as $\mathscr{M}^t_{\alpha(t)} = \langle U, N_U, V^t, \alpha(t) \rangle$, where V^t is a valuation that assigns truth values to every atomic sentence at each world at given time t.

Selecting n at time t, the agent has the neighborhood $B_{\alpha(t)/n}(x^t)$. The truth condition for modal sentences are given by

$$\mathscr{M}^t_{\alpha(t)}, x^t \models \Box p \Leftrightarrow B_{\alpha(t)/n}(x^t) \subseteq ||p||^{\mathscr{M}^t_{\alpha(t)}},$$
$$\mathscr{M}^t_{\alpha(t)}, x^t \models \Diamond p \Leftrightarrow B_{\alpha(t)/n}(x^t) \cap ||p||^{\mathscr{M}^t_{\alpha(t)}} \neq \emptyset.$$

Proposition 31. *1. Every neighborhood is $R_{\alpha(t)}$-definable.*

2. If a neighborhoods $B_{\alpha(t)/n}(x^t)$ satisfies $B_{\alpha(t)/n}(x^t) \subseteq [R_{\alpha(t)}]||p||^{\mathscr{M}^t_{\alpha}}$, then $\mathscr{M}^t_{\alpha(t)}, x^t \models \Box p$ and $\mathscr{M}^t_{\alpha(t)}, x^t \models \Diamond p$.

3. If a neighborhoods $B_{\alpha(t)/n}(x^t)$ satisfies $B_{\alpha(t)/n}(x^t) \cap \langle R_{\alpha}\rangle ||p||^{\mathscr{M}^t_{\alpha}} \neq \emptyset$, then $\mathscr{M}^t_{\alpha}, x^t \models \Diamond p$.

4. If a neighborhoods $B_{\alpha(t)/n}(x^t)$ satisfies $B_{\alpha(t)/n}(x^t) \subseteq [R_{\alpha}](||p||^{\mathscr{M}^t_{\alpha}})^c$, then $\mathscr{M}^t_{\alpha}, x^t \models \neg\Box p$ and $\mathscr{M}^t_{\alpha}, x^t \models \neg\Diamond p$.

5. If a neighborhoods $B_{\alpha(t)/n}(x^t)$ satisfies $B_{\alpha(t)/n}(x^t) \subsetneq [R_{\alpha}]||p||^{\mathscr{M}^t_{\alpha}}$ and $B_{\alpha(t)/n}(x^t) \cap \langle R_{\alpha}\rangle ||p||^{\mathscr{M}^t_{\alpha}} \neq \emptyset$, then $\mathscr{M}^t_{\alpha}, x^t \models \neg\Box p$ and $\mathscr{M}^t_{\alpha}, x^t \models \Diamond p$.

4 Observed Results and Sizes of Neighborhood

In this section, we formulate a way of changing an agent's observations in his neighborhoods. In general, an agent may observe different objects depending on sizes of neighborhoods. We formulate such kind of observations in the dynamic neighborhood model $\mathscr{M}^t_{\alpha(t)}$.

Let us consider two atomic sentences *base*1 and *target*1. $\mathscr{M}^t_{\alpha(t)}, x^t \models base1$ means that a point (world) $x^t \in U$ is in base1. $\mathscr{M}^t_{\alpha(t)}, x^t \models target1$ means that a point $x^t \in U$ is in target1. By observation, an agent can know truth values of modal sentences in $\mathscr{M}^t_{\alpha(t)}$, that is, whether, for example, $\mathscr{M}^t_{\alpha(t)}, x^t \models \Box base1$, $\mathscr{M}^t_{\alpha(t)}, x^t \models \Diamond base1$, and $\mathscr{M}^t_{\alpha(t)}, x^t \models \neg\Diamond base1$ hold or not. In the following, we denote $\mathscr{M}^t_{\alpha(t)}, x^t \models \Box base1$ simply by $\Box base1$ if confusion does not arise. We use a similar notation for other modal sentences.

Observations of Type B Objects

Base1 is a Type B object. When an agent observes $\Box base1$, if he expands the size of his neighborhood, then his observed result changes to $\Diamond base1$. When an agent observes $\neg\Diamond base1$, if he expands the size of his neighborhood, then his observed result changes to $\Diamond base1$. When an agent observes $\Diamond base1$, if he expands the size of his neighborhood, then his observed result changes to $\Box base1$ or $\neg\Diamond base1$ according to positional relationship of the agent and base1. In some cases, observations do not change. Such changes of observations are shown in Fig. 3.

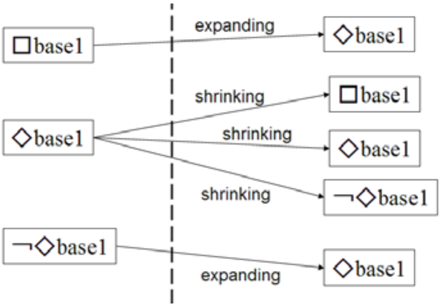


Fig. 3. Changes of observations by expansion and shrinking of neighborhoods for Type B objects

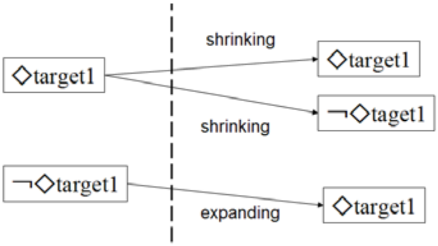


Fig. 4. Changes of observations by expantion and shrinking of neighborhoods for Type A objects

Observations of Type A Objects

Target1 is a Type A object. An agent and a Type A object cannot share the same point at the same time. Let an agent be located at x^t. Then, $\mathscr{M}^t_{\alpha(t)}, x^t \models \neg target1$, and thus $\mathscr{M}^t_{\alpha(t)}, x^t \models \neg\Box target1$. This means that an observed result cannot be $\Box target1$.

When an agent observes $\Diamond target1$, if he shrinks the size of his neighborhood, then his observed result changes to $\neg\Diamond target1$. When an agent observes $\neg\Diamond target1$, if he expands the size of his neighborhood, then his observed result changes to $\Diamond target1$. There are times that the observation will be remain unchanged by expanding or shrinking the neighborhood. Such changes of observations are shown in Fig. 4.

Each Observation and Directions

When an agent observes $\Box base1$, he cannot access or avoid base1 because base1 exists in all directions. When an agent observes $\Diamond base1 \wedge \neg\Box base1$ and gets information of the direction of base1, he can access base1. When an agent observes $\neg\Diamond base1$, he cannot access base1 because he cannot get information of the direction of base1.

In this way, if an agent wants to access or avoid objects, he must select one of neighborhoods that result in $\Diamond object \wedge \neg\Box object$.

5 Simulation

We carried out some simulation using the proposed method. In our simulation, we assume one agent, one target, one base, and walls. The task of the agent is to chase the target. If the agent's energy level is low, he must search the base in order to charge energy.

We perform four kinds of simulation (Fig. 5, 6, 7, 8) where the agent is given different strategies for selecting neighborhoods.

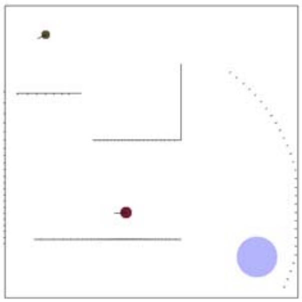


Fig. 5. Agents that select most largest neighborhood

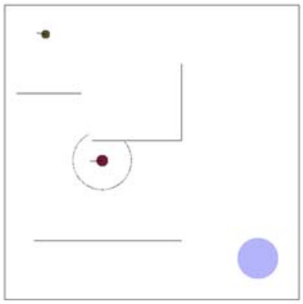


Fig. 6. Agents that shrink its neighborhood randomly

(1) Agents that select most largest neighborhood

If the agent's neighborhood is larger, he tends to avoid more distant walls and thus his mobility becomes more restricted. As a result, his ability of searching the target would be more limited. He can easily find the base, so he seldom runs out of energy.

(2) Agents that shrink its neighborhood randomly

In the shrunk neighborhood, the restriction of agent's mobility is lax because of decrease of number of walls. Despite agent's energy is low, the agent selects small neighborhoods and it cannot find the base and he come to a stop because of running out of energy.

(3) Agents that adopt the neighborhood of the lower approximation of visible region

The agent cannot avoid walls, hence there are no walls in his neighborhood at all times. The agent almost loses the base and he come to a stop.

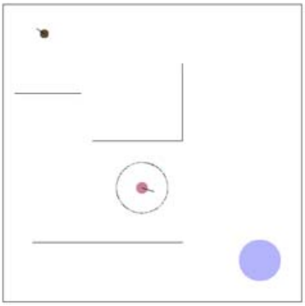


Fig. 7. Agents that adopt the neighborhood of the lower approximation of visible region

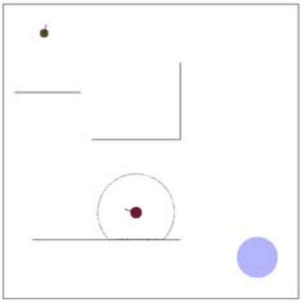


Fig. 8. Agents that adopt the expanded neighborhood of the lower approximation

(4) Agents that adopt the expanded neighborhood of the lower approximation

Even though the agent comes near to walls to a certain degree, he avoids walls.

6 Concluding Remarks

In this paper, we formulated a framework of a variable neighborhood model by a dynamic neighborhood model. We generated rough sets considering an countable fundamental system of neighborhoods and we showed rough sets will be one of strategies for selecting neighborhoods. There is close relationships between the description by rough sets and the description by neighborhood models. We carried out some simulations and confirmed different trends of agents' behaviors by their strategies for selecting neighborhoods. We plan to pursue relationships between agent's behaviors and topological properties. We also plan to consider differences of strategies for selecting neighborhoods.

References

1. Chellas, B.F.: Modal Logic: An Introduction. Cambridge University Press, Cambridge (1980)
2. Pawlak, Z.: Rough Sets. International Journal of Computer and Information Sciences 11, 341–356 (1982)
3. Ubukata, S., Kudo, Y., Murai, T.: An Agent Control Method based on Rough-Set-based Granularity. In: Soft Computing and Intelligent Systems and 9th International Symposium on Advanced Intelligent Systems (SCIS & ISIS 2008), FR-G3-1 (2008)

Counselor, a Data Mining Based Time Estimation for Software Maintenance

Hércules Antonio do Prado[1,2], Edilson Ferneda[1], Nicolas Anquetil[1], and Elizabeth d´Arrochella Teixeira[3]

[1] Catholic University of Brasília, DF, Brazil
[2] EMBRAPA Management and Strategy, Brasília, DF, Brazil
[3] Politec S.A., DF, Brazil
herculesa@{embrapa.br,ucb.br}, eferneda@ucb.br, anquetil@ucb.br, darrochella.ucb@terra.com.br

Abstract. Measuring and estimating are fundamental activities for the success of any project. In the software maintenance realm the lack of maturity, or even a low level of interest in adopting effective maintenance techniques and related metrics, have been pointed out as an important cause for the high costs involved. In this paper data mining techniques are applied to provide a sound estimation for the time required to accomplish a maintenance task. Based on real world data regarding maintenance requests, some regression models are built to predict the time required for each maintenance. Data on the team skill and the maintenance characteristics are mapped into values that predict better time estimations in comparison to the one predicted by the human expert. A particular finding from this research is that the time prediction provided by a human expert works as an inductive bias that improves the overall prediction accuracy.

Keywords: Informal Reasoning, Data Mining, Software Maintenance.

1 Introduction

Both the increasing technological evolution and the speed in which software requirements change have led the software engineering community to consider software as an object in permanent improvement. Consequently, a great demand for maintenance has settled down. On the other hand, the high competitive scenery for software services suppliers stimulates software companies to seek alternatives to improve their configuration management in order to provide the best service for their clients. Adopting more effective methods for the development process and for maintenance practices is mandatory to reach that objective (LEHMAN, PERRY & RAMIL, 1998).

The concerns about the time applied to maintain systems appeared by the nineties. Soon, the community becomes aware that by 90% of the total cost of a system is devoted to its maintenance (PIGOSKI, 1996; PFLEEGER, 2001). In general, the research on software metrics and estimation have received increasing attention in the Software Engineering (SE) context, looking for satisfying the clients requirements of quality, quickness, convenience, and low cost for implementation maintenance. Such

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 364–371, 2009.

methods have pushed the development of high quality software products with adequate costs.

This paper presents the results of a research focused on predicting maintenance time based in variables like platform, kind of service, complexity, and service importance. An approach, named *Counselor*, is proposed to support an analyst in deciding about a schedule to carry out a maintenance task. This approach includes a set of regression models that can be applied alternatively as an optimistic, a pessimistic, or a middle way predictor.

2 The Time Estimation in Software Engineering

Measuring is the process by which numbers and symbols are assigned to the real world in order to enable characterizing each entity by means of clearly defined rules (FENTON and PFLEEGER, 1997). According to Pressman (2009), measures can be defined as quantitative indicators to express the extension, dimension, capacity, or size of an attribute of a product or process.

Also, according to Fenton e Pfleeger (1997), people measure to understand, control, and improve the software development process. As stated by Brito e Abreu (1992), metrics are not the panacea for all problems in the software development process. However, if used conjointly with other concerns, like a careful recruitment of the development team and a practice of quality control, they have been pointed out as a catalyst for quality and productivity.

The main results expected when applying some metric program are (BRITO E ABREU, 1992): *(i)* improving the management activity, mitigating the uncertainty related to the software development; *(ii)* a more effective quality control, *(iii)* anticipating problems, allowing corrective actions; *(iv)* improving the enterprise reputation; and *(v)* higher productivity, among others.

Although the great amount of available literature related to software estimation, many barriers preclude the developers of adopting such techniques. Facts like low maturity, lack of interest of development enterprises, and low acceptance can be enlisted as the main causes for failure in adopting metrics. And, as a consequence of having the measuring process not well rooted in the enterprises, we have the high maintenance cost of information systems that the research community observes.

There exists a bunch of metrics, almost of them able to measure systems in many phases, before, during, and after the development. However, important issues like system size, development language, applied databases, team experience, and so on, are not considered in the existing metrics. In the lack of better options for maintenance time prediction, many institutions adapt measuring techniques for software size, satisfying precariously their necessities.

3 Material and Methods

Counselor includes a set of regression models that can be applied under the optimistic, pessimistic, or middle way adopted approaches. The regression models were built according the CRISP-DM method for Knowledge Discovery in Databases (CHAPMAN, 2000) and combined in a committee machine adapted to regression.

3.1 Data Description

The variables considered to build the regression models were defined according the adherence they have with the related task. This adherence was specified by software engineering practitioners and reflects the importance the variables have to predict the expected maintenance time. Also, the availability of historical observations was considered to define the independent variables.

A set of 21,493 records of software maintenance from a Brazilian software company were taken for creating the regression models. This data set was used both for training and testing the models by applying *n*-fold cross-validation approach. Each model was built from the same data set, varying only the learning algorithm. The variables used as input to the model were: main platform, kind of service, subtask involved, team experience, data base, system complexity, the time estimation from an expert, and how critical the system is. During the pre-processing activity, records with no information in some field or wrong data type were considered outliers and taken off from the original data set. The resulting data set contains 18,845 records, distributed according to the platform are shown in Table 1.

Table 1. Amount of records by platform

Platform	# Records
Three Layers	35
Asp-Sql-server-VB	29
Delphi-Oracle	96
Client Server	659
SAP	1,154
Mainframe	6,435
Web	10,437
Total of records	18,845

3.2 Modeling the Problem

Regression is a powerful data mining approach and is used to investigate the relation among variables. It is a data analysis technique that builds models in which the objective is to predict numeric values. A regression model can adjust a function over the data, minimizing the error from the prediction values against the observed values. In other words, statistically, regression is a method to estimate a conditional (expected value) from a variable *y*, given the values from some other variables. Linear regression is the vanilla approach for regression and assumes that the relation among the independent and the conditional variables presents a linear behavior, under some parameters. Otherwise, the regression is called non-linear.

The methods we have chosen to work with are described next, include both linear and non-linear approaches, and have as an important characteristic the ability to work most naturally with nominal attributes at the input space. They are all implemented in Weka benchmark (WITTEN and FRANK, 2005).

- *LeastMedSq*. It is an implementation of a least median squared linear regression that minimizes the median squared error.
- *Linear Regression*. Although requiring the strong restriction of linearity on the model, the linear regression was used in this work as a bottom line model.
- *Multilayer Perceptron*. Implements a back propagation learning method.
- *RBF*. The radial basis function neural network has as main characteristic the use of a normalized distance (based in a Gaussian function) between the input points and the hidden nodes to define the activation of each node. The activation is as stronger as closer are the two points.
- *SMOreg*. This is an implementation of the sequential minimal optimization algorithm for training a support vector regression model. Its main characteristics are to cope with missing values and to transform nominal attributes into binary ones.

For each platform, it was generated twenty models, the five types above mentioned combined with two validation schemas (2- and 10-fold cross validation) and two variations in the input space. These variations in the input space refer to the presence or absence of the *expert time prediction* as independent variable. Considering that we studied seven platforms, the amount of models generated reaches hundred forty models.

4 Results and Discussion

In order to provide a basis for comparison among the resulting regression models we adopted the Mean Absolute Error (MAE) as accuracy evaluator. The estimated time for each maintenance request came from the feeling of experimented software engineering practitioners, with large knowledge in each platform. On the other hand, the time observed for each request was taken as the elapsed time between the beginning and the maintenance completion.

Also, it was observed that the models generated using the *expert time prediction* as independent variables usually have shown a smaller error in the prediction than without this variable. It can be interpreted that the *expert time prediction* variable works as a positive bias, improving the overall prediction performance. If we do not consider *expert time prediction* as input variable, the best model generated came from the LeastMedSq algorithm, for the platforms: 3-Layers, Asp, and Delphi-Oracle and the SMOreg for Client-Server, Mainframe, SAP, and Web. When including *expert time prediction* in the input space, SMOreg generated the best model for the platforms 3-layers, Asp, Delphi-Oracle, and Mainframe and the LeastMedSeq for Client-Server and SAP. With respect to the validation schemas, it was not observed significant differences when using 2- or 10-fold cross validation. So, in this study we show only the five original models with *expert time prediction* as independent variable, tested under the 2-fold cross validation schema.

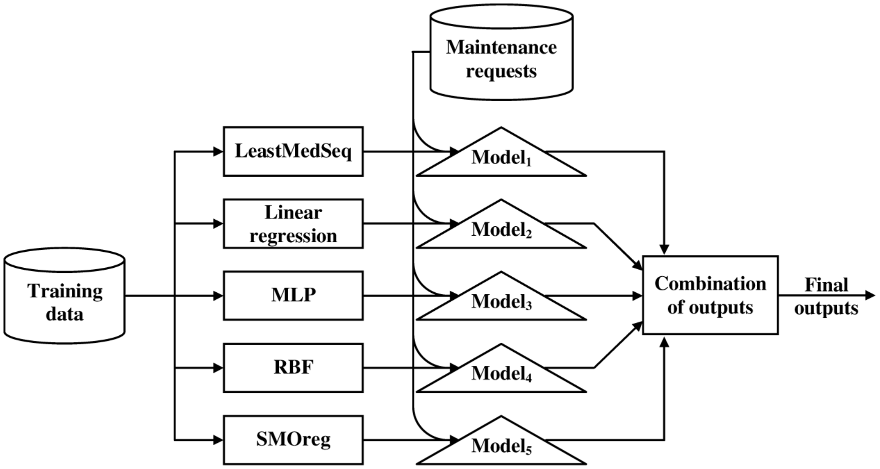


Fig. 1. Committee machine for combining the regression models

It was observed that the time predicted by the models is, in general, smaller than the expert prediction. It can be seen from Table 2 that the time adopted by the enterprise was always bigger than the models predictions. However, it does not mean that one must take the smallest predicted time. Considering the risk of losses by taking a very small prediction that, at the end of the task, will not be verified, we propose three approaches, one pessimistic, one optimistic, and a middle way approach. In the pessimistic approach, the analyst takes the value from the model with bigger MAE. For the optimistic approach, the model with smaller MAE is considered. Finally, for the middle way approach, we propose a committee machine (HAYKIN, 2008) that generates a output a combination of the outputs from the five models (Fig. 1).

Table 2. MAE for each model/platform, with and without the expert prediction

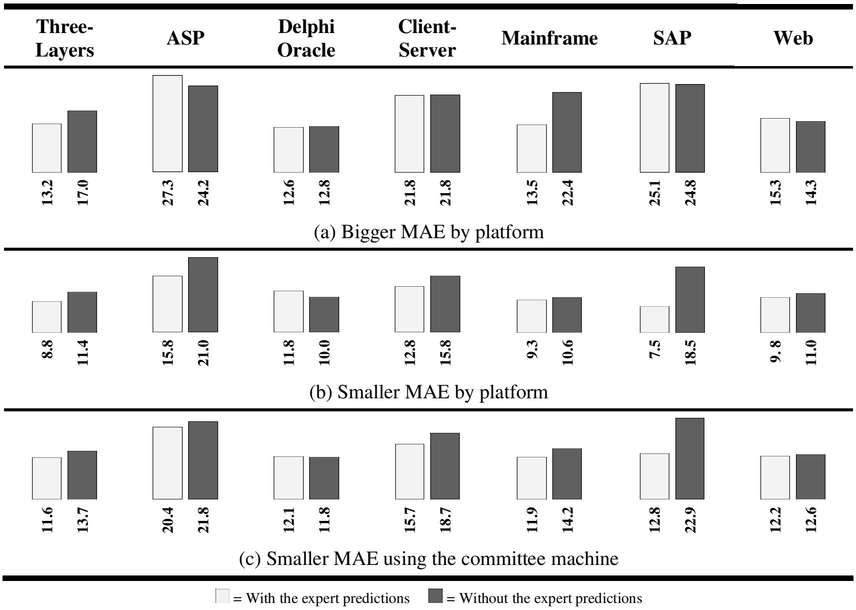
Figures		Three-Layers	Asp	Delphi Oracle	Client-Server	Mainframe	SAP	Web
Amount or records		35	29	96	659	6435	1,154	10,437
Error from the expert		22.4981	14.2880	26.4196	40.8153	65.7706	36.7400	59.4310
LeastMedSeq	X	10.8377	17.3384	11.9762	**12.8089**	11.0882	**7.5221**	10.9888
	W	***11.4304***	***20.9931***	***10.0356***	17.1368	10.9751	19.0501	11.3921
Linear Regression	X	11.1096	18.9912	12.5058	14.4096	11.9036	11.2049	11.8894
	W	14.2390	21.448	12.5058	19.0364	13.3773	24.4608	14.3395
Multilayer Perceptron	X	13.2489	27.3497	11.8749	16.2361	13.5023	12.5069	12.9170
	W	17.0435	24.1743	12.2051	19.6620	22.3555	27.7545	13.1250
RBF	X	13.8710	22.4911	12.5628	21.8305	13.6017	25.1537	15.2921
	W	14.0309	21.3689	12.8124	21.8333	13.8446	24.7638	13.3096
SMOreg	X	**8.7880**	**15.7709**	**11.7955**	13.0251	**9.3207**	7.7925	**9.7835**
	W	11.8528	21.2248	11.6718	***15.7621***	***10.6507***	***18.511***	***11.004***

X = With expert prediction. **W** = Without expert prediction.

In the example shown, the result generated by the committee machine is the arithmetic mean of the five outputs. Other variations, like the weighted mean could be adopted.

Table 2 shows the values estimated by each model, under the 2-fold cross validation schema, for each platform. The best prediction is marked in bold face, considering the option: using or not using (in italic) the expert prediction as independent variable.

The MAE for each model is depicted in Figure 2. The bigger MAE, related to the pessimistic approach, are shown in Figure 2a while the optimistic approach (smaller MAE) are illustrated in Figure 2b that shows the best results. In Figure 2c it is shown the arithmetic mean among the MAE from the committee machine. The relations between the expert predictions and the committee machine predictions become clear in this figure.



(a) Bigger MAE by platform

(b) Smaller MAE by platform

(c) Smaller MAE using the committee machine

Fig. 2. MAE from the models with and without the expert predictions

A comparison of expert predictions with models with bigger and smaller MAE, with or without the expert prediction as independent variable, and the committee machine is show in Figure 3. With the only exception in ASP platform, all MAE from the models are smaller than the MAE of expert predictions.

For illustrating the possible use of *Counselor*, we took three maintenance requests, with the time consumed for their completion, and applied the models for prediction. The results can be seen at Table 3.

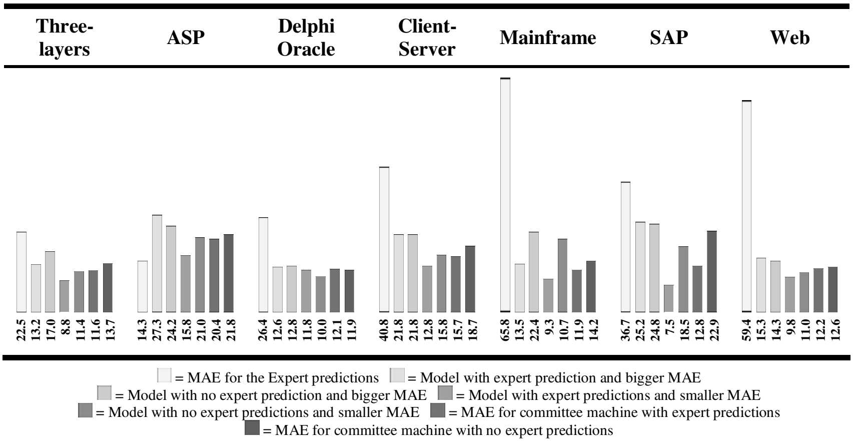


Fig. 3. Comparing MAE of expert prediction with models and committee machine

Table 3. Models (with expert predictions) applied for three maintenance cases for Mainframe

#	Time consumed	Expert prediction	SMOreg (optimistic)	Mean (middle way)	RBF (pessimistic)	Least MedSq	Linear regresion	Multilayer Perceptron
1	4	2	2.754	* 3.773	6.552	1.079	4.709	-3.764
2	6	6	* 7.017	9.489	16.684	5.699	8.557	-6.295
3	12	12	4.341	* 7.7122	16.677	9.831	-12.387	-9.210

Predictions from the Multilayer Perceptron were completely discarded since the model predicted only negative values, which constitute spurious values. The optimistic predictions (from SMOreg) are shown in the fourth column, the middle way in the fifth, and the pessimistic in the sixth. Comparing these predictions with the ones issued by the expert (third column), it can be observed that the three predictions (optimist, pessimistic, and middleway) can provide some figures to help the expert in deciding if his predictions are too far from the historical set of requests. For example, in request 1, *Counselor* best prediction was 3.7 hours, that is better than the expert prediction. The predictions nearer the amount of time consumed at the maintenance completion, for each approach, are marked with "*".

5 Conclusion

We have shown, empirically, that we can estimate with an acceptable precision the time necessary to carry out a software maintenance task. Considering the increasing competition in the modern economy, in all segments, a tool like *Counselor* can help a software development company in defining adequately their prices in the market. Such a model can help the manager in offering feasible prices for the services while assuring the necessary quality of these services. This approach can be improved permanently with new maintenance requests and with the expansion of expert experience.

References

Brito e Abreu, F.: Metrics in Project Management of Information Systems Development. Proceedings of the 6^{a} Jornada de Qualidade. APQ, Lisbon (1992) (in Portuguese)

Chapman, P., et al.: CRISP-DM 1.0 - Step-by-step Data Mining Guide (2000), `http://www.crisp-dm.org/CRISPWP-0800.pdf`

Fenton, N., Pfleeger, S.: Software Metrics - A Rigorous & Pratical Approach, 2nd edn. PWS Publishing Company, Boston (1998)

Haykin, S.: Neural Networks: A Comprehensive Foundation, 3rd edn. Prentice Hall, New Jersey (2008)

Lehman, M.M., Perry, D.E., Ramil, J.F.: Implications of Evolution Metrics on Software Maintenance. In: Proceedings of the international Conference on Software Maintenance (ICSM). IEEE Computer Society, Washington (1998)

Pfleeger, S.L.: Software Engineering: Theory and Practice, 2nd edn. Prentice Hall, New Jersey (2001)

Pigoski, T.M.: Practical Software Maintenance. John Wiley & Sons, Inc., New York (1996)

Pressman, R.: Software Engineering: A Practitioner's Approach, 7th edn. McGraw-Hill Higher Education, New York (2009)

Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools and techniques, 2nd edn. Morgan Kaufmann, San Francisco (2005)

An Integrated Knowledge Adaption Framework for Case-Based Reasoning Systems

Ning Lu, Jie Lu, and Guangquan Zhang

Faculty of Engineering and Information Technology
University of Technology Sydney, PO Box 123, Broadway, NSW 2007, Australia
{philiplu,jielu,zhangg}@it.uts.edu.au

Abstract. The development of effective knowledge adaption techniques is one of the promising solutions to improve the performance of case-based reasoning (CBR) systems. Case-base maintenance becomes a powerful method to refine knowledge in CBR systems. This paper proposes an integrated knowledge adaption framework for CBR systems which contains a meta database component and a maintenance strategies component. The meta database component can help track changes of interested concepts and therefore enable a CBR system to signal a need for maintenance or to invoke adaption on its own. The maintenance strategies component can perform cross-container maintenance operations in a CBR system. This paper also illustrates how the proposed integrated knowledge adaption framework assists decision makers to build dynamic prediction and decision capabilities.

Keywords: knowledge adaption, case-based reasoning, case-base maintenance, machine learning, adaptive decision support.

1 Introduction

Organizational decision makers are expected to carry out tasks and solve problems effectively and efficiently by learning previous knowledge. Machine learning provides automatic techniques to learn past knowledge, observations, and various experience in order to make accurate predictions and support decision-making. A main concern in machine learning is the design and development of learning algorithms that can continuously acquire and refine knowledge, such as rules and patterns, from data, and allow computers to improve their performance over time.

One of the current challenges facing machine learning algorithms, like CART, ID3, C4.5 and IFN [1], is to handle the dynamic changing situations. In many machine learning-based systems, the distributions or patterns underlying the data are changing over time rather than remaining stable, which is also known as concept drift [2]. As a result, when a learning algorithm considers all the past training examples or makes an assumption that training data is a random sample drawn from a stationary distribution, the induced patterns may not relevant to the new data [3, 4]. This raises an urgent need of an effective learner which is able to track such changes and to quickly adapt to them [2].

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 372–379, 2009.

Case-based reasoning (CBR), as one of the most popular machine learning technique, shows the process of solving new problems based on the solutions of similar past problems [5]. The prediction and reasoning accuracy is mainly depended on the quality of cases and other knowledge. In order to perform as well as possible despite changing circumstances, a CBR system must be able to signal the need for maintenance or to invoke maintenance strategies as needed, that is to develop an adaptive CBR system [6]. This leads to a new active research area, case-based maintenance (CBM), which has been seen as a new phase of CBR. Literature reports several CBM policies to improve efficiency, reserve competency and maintain consistency of a CBR system. The most creative point of the CBM technique is that it allows a CBR system to update itself dynamically by taking information from problem solving process. Obviously, this ability can greatly improve the performance of CBR in applications. We therefore can use CBM techniques to support the establishment of adaptive CBR systems.

The success of long-term CBR applications in dynamic environment depends critically on the accuracy and refining of the contents of case bases and other knowledge containers — vocabulary knowledge, retrieval knowledge (similarity measure), and adaption knowledge (solution transformation) [7]. To provide maintenance support for a case base, Smyth and Keane [8] suggested an competence-preserving case deletion policy which introduces two important competence properties, coverage and reachability. Delany & Cunningham [9] enhanced this competence model by adding a liability property and proposed a conservative redundancy reduction policy. In the meantime, Zhu & Yang [10] developed a case-addition policy, which repeatedly selects cases from an original case-base and then adds them into an empty case-base until reaches a certain size limit. Moreover, some researches of CBM extended beyond maintaining case base alone and addressed to all the other knowledge containers of a CBR system. Kim & Han [11] described a GA-based case representation which can search for the near-optimal form of representation through discretization; Craw, Jarmulak & Rowe [12] developed a method to optimize CBR retrieval and Hanney & Keane [13] presented an inductive learning algorithm to extract adaption rules from the cases in the case-base which is very useful in CBM. All these proposed methods can maintain case base and other knowledge, but they mainly focus on single knowledge container only. In real world applications, a completed CBR system contains a set of knowledge containers and must be able to adapt knowledge within all these containers as a whole. Heister and Wilke [14] identified this problem and suggested a repair phase to keep consistency among knowledge containers as well as a framework to support the new phase. However, their framework is weak in tracking changes and lack of discussion of relationships between adaption knowledge and other knowledge in a CBR system. To handle these two issues, we propose an integrated knowledge adaption framework, which has a meta database component to help track changes of concepts thus enable the CBR system to signal a need for maintenance and a maintenance strategies component for cross-container adaption operations.

The paper is organized as follow. In Section 2, we briefly explain related researches and developments in CBR and CBM. In Section 3, we present an integrated knowledge adaption framework for CBR systems. Comparing with others, the advantages of the framework is it gives detailed approaches to enhance the communication between a CBM/CBR system and its users and to restore all knowledge containers dynamically. Section 4 gives an example to illustrate how the framework design commits a cross-container adaption for CBR. Conclusions and further studies are discussed in Section 5.

2 Related Works

This section will first review the concepts and methods of case-based reasoning. It will then introduce the new development in CBR research by considering CBM.

2.1 Case-Based Reasoning

Case-based reasoning offers a powerful learning ability to use past experiences as a basis for dealing with new problems, and facilitates the knowledge acquisition process by avoiding the time required to elicit solutions from experts. It is represented by a four-step (4Rs) cycle: *retrieve, reuse, revise* and *retain* [5]. A new case defined by an initial problem description is used to *retrieve* a case from the base of previous cases. "The retrieved case is combined with the new case - through *reuse* - into a solved case, i.e. a proposed solution to the initial problem. Through the *revise* process this solution is tested for success, e.g. by being applied to the real world environment or evaluated by a teacher, and repaired if failed. During *retain*, useful experience is retained for future reuse, and the case base is updated by a new learned case, or by modification of some existing cases."[5]

2.2 Extended Case-Based Reasoning

The success of a CBR system is mainly dependent on the suitability of knowledge and the correctness of reasoning. Therefore, certain "harmful" knowledge may actually degrade system performance [8]. To cope with all issues that arise when CBR is used in real world applications and especially in rapid changing environments, some additional phases such as maintenance phase has been accepted to extend original CBR cycle [15]. As a typical development, Reinartz [16] extended the standard four-step CBR cycle [5] by two additional steps: *review* and *restore*. The *review* step covers tasks to judge and monitor the current state of a CBR system and its knowledge containers, whereas the *restore* step invokes mechanisms to change the system and its knowledge to improve the performance. Figure 1 shows the six-step CBR cycle.

2.3 Case-Base Maintenance

The case-base maintenance (CBM) refers to the process of refining a CBR system to improve the system's performance, which implements policies of revising the contents of a CBR system in order to facilitate future reasoning [17]. To guide the general study of maintenance in a CBR system, Wilson and Leake [17] provided a common framework for describing CBM policies. They categorized CBM approaches in terms of CBM policies that determine when and how a CBR system performs CBM. But they only gave a very general framework, which didn't give enough design and explanation about how operations of different knowledge containers affect each other and how to integrate maintenance operations of different containers into a CBR system.

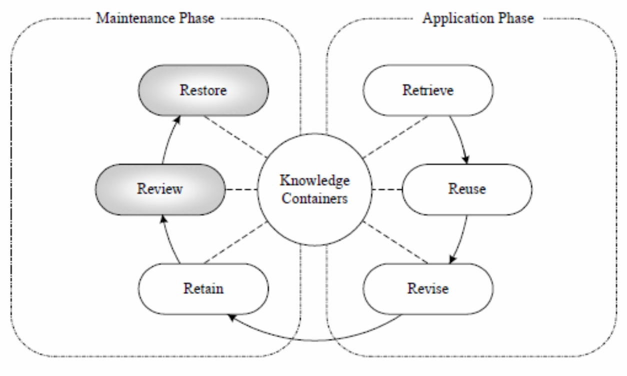


Fig. 1. The Six-Step CBR Cycle [16]

To address this issue, Heister and Wilke [14] presented a two-phase approach. They defined maintenance operations as modifications of a system made by the users in an intended manner and repair operations as pre-defined scripts to be performed after a maintenance operation has happened to keep system consistent. They further developed a maintenance architecture to support these two operations during changes of the system. However, their framework is not able to track changes of concepts, and thus doesn't facilitate the decisions of when, where and how to make changes.

3 An Integrated Knowledge Adaption Framework for CBR

To provide better communications with the decision makers and facilitate their decisions with regards to knowledge maintenance issue, we proposed an integrated knowledge adaption framework for CBR systems. The framework is able to track changes of concepts, thus signal a need for maintenance to decision makers or even perform a series of operations on its own. It also supports changes of multiple knowledge containers and cross-container operations. Coincide with the extended six-step of CBR cycle [16], the framework consists of two parts, application phase (the right half) and maintenance phase (the left part), as shown in Figure 2. This paper mainly explains the *meta database* component (corresponding to the *review* step in Figure 2) and *maintenance strategies* component (corresponding to the *restore* step in Figure 2).

3.1 Meta Database

Followed by the general study of maintenance policies [17], we defined a meta database which stores the information gathered about individual cases, about the case base as a whole or in part, and about the overall processing behaviour of the CBR system. The information gathered at individual case level includes the number of times that a case has been reused successfully or the times it has failed, the histories of case retention or reuse and the current measurement of the coverage set and reachability set for a case [18]. The information gathered at case base level includes cases in groups

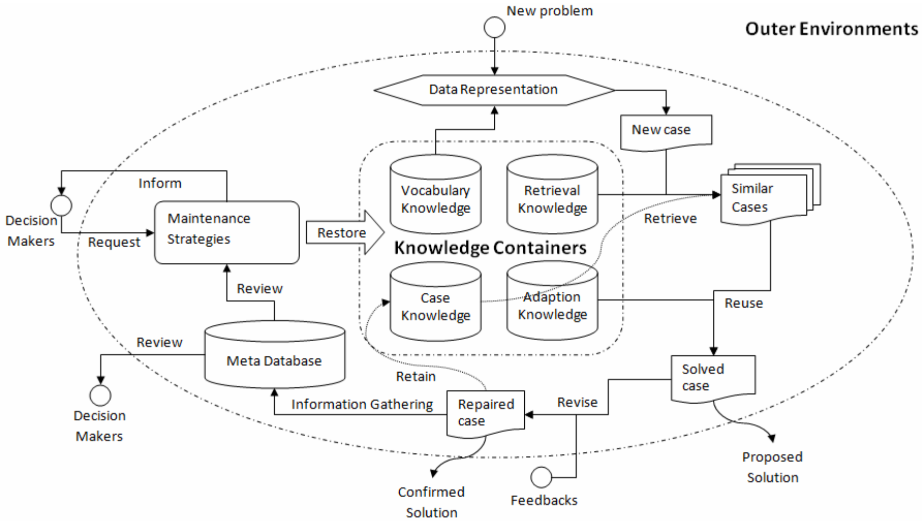


Fig. 2. Integrated knowledge Adaption Framework for Case-based Reasoning System

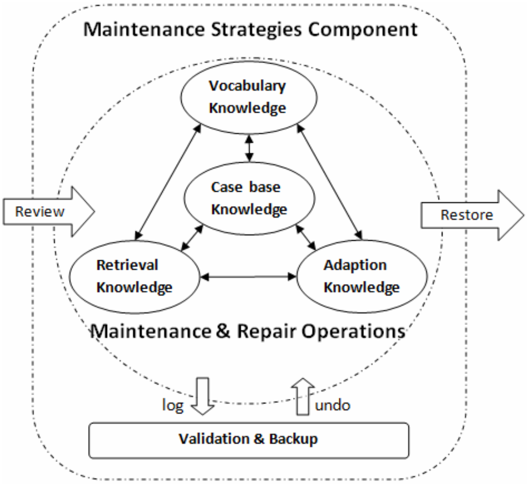


Fig. 3. Maintenance Strategies Component

which support the generalized adaption knowledge or trained similarity measures and the size of current case base. The information gathered for processing behaviour includes the average retrieval time of the system, a set of problems that the system cannot solve successfully or cost highly for solving. By reviewing or monitoring the information gathered in it, one could assess the current performance of a system, thus perform or trigger a certain set of maintenance operations when desired.

3.2 Maintenance Strategies

The result of data analysis serves as input for determining whether CBM is necessary and where to restore. Thus, the maintenance strategies component takes *meta data* as input and restores the knowledge in a CBR system. While a system may inform users that certain maintenance operations are needed, such as removing noise cases, refining similarity measures, users may also request certain operations in desired, such as adding new adaption rules, adding new similarity measures. As changes of one knowledge container may also affect others, the proposed or requested maintenance operations will be followed by certain sets of repair operations to prohibit inconsistency as shown in Figure 3. The results of these operations will be evaluated by the performance improvement of a CBR system or the decision makers, and the system might be reverted in case of failure. We will give an example in the next section to illustrate how the changes will be made.

4 A Case Study

Suppose there is a CBR system designed for a telecommunication company, which helps the marketing department to determine strategies on how to attract more customers. In the past, to attract the youth, several cases indicated that a low charge of SMS was a good champion strategy (depicted as circles in Figure 4). Recently, as iphone becomes a new hot spot and the functionality of playing multimedia is more

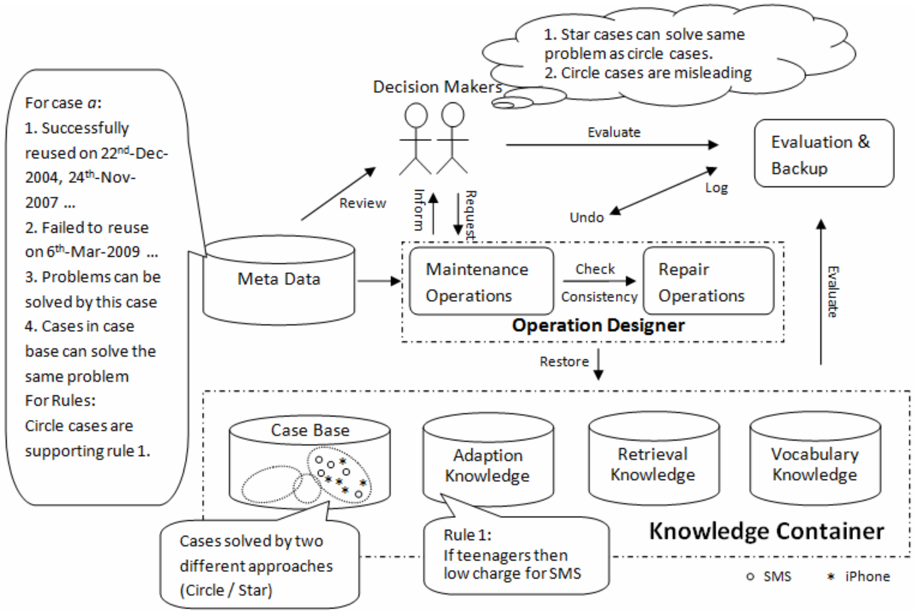

1. Star cases can solve same problem as circle cases.
2. Circle cases are misleading
Decision Makers
For case *a*:
1. Successfully reused on 22nd-Dec-2004, 24th-Nov-2007 ...
2. Failed to reuse on 6th-Mar-2009 ...
3. Problems can be solved by this case
4. Cases in case base can solve the same problem
For Rules:
Circle cases are supporting rule 1.
Evaluate
Evaluation & Backup
Review
Inform
Request
Log
Undo
Meta Data
Maintenance Operations
Check
Consistency
Repair Operations
Operation Designer
Evaluate
Restore
Case Base
Adaption Knowledge
Retrieval Knowledge
Vocabulary Knowledge
Knowledge Container
Cases solved by two different approaches (Circle / Star)
Rule 1:
If teenagers then low charge for SMS
○ SMS
* iPhone

Fig. 4. An Example of Case-base Maintenance under the Integrated Knowledge Adaption Framework

attractive, the original strategy does not work well in the competition. To solve this problem, new solutions are worked out and retained as new cases in the case base (stars in Figure 4), such as offering a good price for iphone. However, when marketing decision-makers consider both scenarios, they will get confused, as the two strategies (expressed by two groups of cases) are conflict with each other. Thus case base maintenance will be needed.

As shown in Figure 4, by reviewing *meta data*, the decision makers are able to recognize increasing inconsistency for solving the problem of how to attract the youth and a decreasing reliability of circle cases. Thus removing of such old fashioned cases will certainly improve the performance of the CBR system. This proposed system can also automatically identify these 'old' cases and inform/perform a case deletion operation through suitable review and evaluations.

In addition, changes in a case base may affect other knowledge in the CBR system, especially when the original retrieval and adaption knowledge is obtained based on the case base. Take the same example, as the result of the removal of circle cases in the case base, the support of generalized adaption knowledge — Rule 1 (Figure 4), becomes weak. Thus certain repair operations may need to be performed to keep system consistent. For example, Rule 1 will be replaced by new rules generated by the remaining cases in the case-base.

5 Conclusions and Future Works

Traditional machine learning methods are facing challenges when applied in dynamic changing environments. For CBR systems, CBM has already been accepted as an essential phase. In this paper, we have proposed an integrated knowledge adaption framework which integrates all knowledge containers together and supports cross-container operations. Especially, this framework adopted a meta database component which gathers information from three different aspects and helps to determinate when and where to restore a CBR system.

For future work, noise cases always mislead the adaption process and can hardly be distinguished with newly concepts. Maintenance strategies under different environments still need further studies and investigations. Second, maintenance strategies themselves could also be seen as knowledge in a CBR system. Studies on how to assess and refine this kind knowledge will be another task.

Acknowledgment. The work presented in this paper was supported by Australian Research Council (ARC) under Discovery Project DP0880739.

References

1. Maimon, O., Last, M.: Knowledge Discovery and Data Mining - The Info-Fuzzy Network (IFN) Methodology 2000. Kluwer Academic, Boston (2000)
2. Widmer, G., Kubat, M.: Learning in the Presence of Concept Drift and Hidden Contexts. Machine Learning 23(1), 69–101 (1996)

3. Geoff, H., Laurie, S., Pedro, D.: Mining time-changing data streams. In: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, San Francisco (2001)
4. Cohen, L., et al.: Info-fuzzy algorithms for mining dynamic data streams. Applied Soft Computing 8(4), 1283–1294 (2008)
5. Aamodt, A., Plaza, E.: Case-based reasoning: Foundational issues, methodological variations, and system approaches. AI Communications 7(1), 39–59 (1994)
6. Leake, D., Wilson, D.: When Experience is Wrong: Examining CBR for Changing Tasks and Environments. In: Case-Based Reasoning Research and Development, p. 720 (1999)
7. Richter, M.M.: Introduction. In: Lenz, M., et al. (eds.) Case-Based Reasoning Technology. LNCS (LNAI), vol. 1400, pp. 1–15. Springer, Heidelberg (1998)
8. Smyth, B., Keane, M.T.: Remembering To Forget: A Competence-Preserving Case Deletion Policy for Case-Based Reasoning Systems. In: Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence. Morgan Kaufmann, San Francisco (1995)
9. Delany, S.J., Cunningham, P.: An Analysis of Case-Base Editing in a Spam Filtering System. In: Advances in Case-Based Reasoning, pp. 128–141 (2004)
10. Zhu, J., Yang, Q.: Remembering to Add: Competence-preserving Case-Addition Policies for Case-Base Maintenance. In: Proceedings of the International Joint Conference in Artificial Intelligence (IJCAI). Morgan Kaufmann, San Francisco (1999)
11. Kim, K.-j., Han, I.: Maintaining case-based reasoning systems using a genetic algorithms approach. Expert Systems with Applications 21(3), 139–145 (2001)
12. Craw, S., Jarmulak, J., Rowe, R.: Maintaining Retrieval Knowledge in a Case-Based Reasoning System. Computational Intelligence 17(2), 346 (2001)
13. Hanney, K., Keane, M.: Learning adaptation rules from a case-base. In: Advances in Case-Based Reasoning, pp. 179–192 (1996)
14. Heister, F., Wilke, W.: An Architecture for Maintaining Case-Based Reasoning Systems. In: Advances in Case-Based Reasoning, p. 221 (1998)
15. Göker, M.H., Roth-Berghofer, T.: The development and utilization of the case-based helpdesk support system HOMER. Engineering Applications of Artificial Intelligence 12(6), 665–680 (1999)
16. Reinartz, T., Iglezakis, I., Roth-Berghofer, T.: Review and Restore for Case-Base Maintenance. Computational Intelligence 17(2), 214 (2001)
17. Wilson, D.C., Leake, D.B.: Maintaining Case-Based Reasoners: Dimensions and Directions. Computational Intelligence 17(2), 196 (2001)
18. McKenna, E., Smyth, B.: Competence-Guided Case-Base Editing Techniques. In: Advances in Case-Based Reasoning, pp. 235–257 (2000)

A Logical Anticipatory System of Before-After Relation Based on Bf-EVALPSN

Kazumi Nakamatsu[1], Jair Minoro Abe[2], and Seiki Akama[3]

[1] University of Hyogo, Himeji, Japan
nakamatu@shse.u-hyogo.ac.jp
[2] Paulista University, Sao Paulo, Brazil
jairabe@uol.com.br
[3] University of Tsukuba, Tsukuba, Japan
sub-akama@jcom.home.ne.jp

Abstract. A paraconsistent annotated logic program called bf-EVALPSN has been developed for dealing with before-after relations between processes and applied to real-time process order control. In this paper, we propose a logical aticipatory system for before-after relation between processes based on reasoning of bf-EVALP vector annotations.

Keywords: before-after relation, EVALPSN, bf-EVALPSN, annotated logic program, anticipatory system.

1 Introduction

We have already developed a paraconsistent annotated logic program called Extended Vector Annotated Logic Program with Strong Negation(abbr. EVALPSN), which can deal with intelligent control and safety verification such as pipeline process control [1,2,3]. We have also developed an EVALPSN called bf(before-after)-EVALPSN to deal with bf(before-after)-relations between time intervals (processes), and applied it to real-time process order control. In bf-EVALPSN, a particular EVALPSN literal $R(p_i, p_j, t) : [(i, j), \mu]$ is introduced and its vector annotation (i, j) is used for representing the bf-relation between processes pi and pj according to process start/finish time information. We propose a real-time computing system for bf-relations between two processes in bf-EVALPSN.

Suppose that we deal with n processes and their bf-relations in bf-EVALPSN, then ${}_n\mathrm{C}_2$ bf-relations should be computed. However, it is not so efficient to deal with all ${}_n\mathrm{C}_2$ bf-relations independently, therefore, we also propose a real-time bf-relation reasoning system which can reason the vector annotation of $R(p_0, p_2, t)$ from those of $R(p_0, p_1, t)$ and $R(p_1, p_2, t)$ in real-time.

2 Before-After EVALPSN

First of all, we introduce a particular literal $R(p_i, p_j, t)$ whose vector annotation represents the before-after relation between processes Pr_i and Pr_j at time t, and it is called a *bf-literal*[1] [5,6].

[1] Hereafter, the word "**b**efore-a**f**ter" is abbreviated as just "bf" in this paper.

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 380–387, 2009.

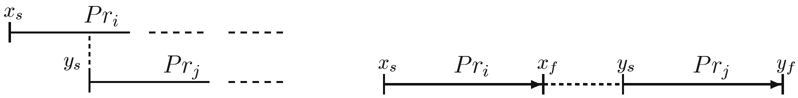


Fig. 1. Before/After and Disjoint Before/After

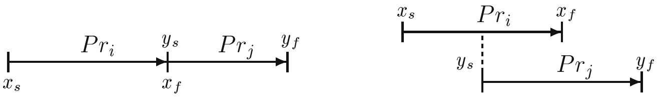


Fig. 2. Immediate Before/After and Joint Before/After

Definition 1 (bf-EVALPSN)
An extended vector annotated literal, $R(p_i, p_j, t) : [(i, j), \mu]$ is called a *bf-EVALP literal*, where (i, j) is a vector annotation and $\mu \in \{\alpha, \beta, \gamma\}$. If an EVALPSN clause contains bf-EVALP literals, it is called a *bf-EVALPSN clause* or just a *bf-EVALP clause* if it contains no strong negation. A *bf-EVALPSN* is a finite set of bf-EVALPSN clauses.

We provide a paraconsistent before-after interpretation for vector annotations of bf-literals, and such a vector annotation is called a *bf-annotations*. Exactly speaking, bf-relations between processes are classified into meaningful fifteen kinds according to bf-relations between start/finish times of two processes in bf-EVALPSN. We briefly review bf-EVALPSN [4].

Before (be)/After (af). First of all, we define the most basic bf-relations *before/after* as one of two processes starts before another one, which are represented by bf-annotations be$(0, 8)$/af$(8, 0)$. They are described as the left process time chart in Figure 1. We define other 13 kinds of bf-annotations as well as before(be)/after(af).

Disjoint Before (db)/After (da). bf-relations *disjoint before/after* are represented by bf-annotations db$(0, 12)$/ da$(12, 0)$ and described as the right process time chart in Figure 1.

Immediate Before (mb)/After (ma). bf-relations *immediate before/after* are represented by bf-annotations mb$(1, 11)$/ ma$(11, 1)$ and described as the left process time chart in Figure 2.

Joint Before (jb)/After (ja). bf-relations *joint before/after* are represented by bf-annotations jb$(2, 10)$/ja $(10, 2)$ and described as the right process time chart in Figure 2.

S-included Before (sb)/After (sa). bf-relations *s-included before/after* are represented by bf-annotations sb$(3, 9)$/sa $(9, 3)$ and described as the left process time chart in Figure 2.

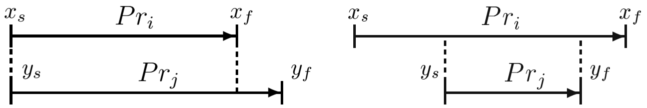


Fig. 3. S-included Before/After, and Included Before/After

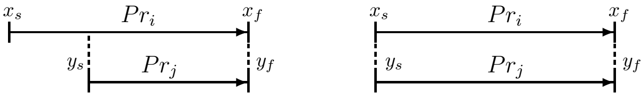


Fig. 4. F-included Before/After and Paraconsistent Before/After

Included Before (ib)/After (ia). bf-relations *icluded before/after* are represented by bf-annotations ib$(4, 8)$/ia $(8, 4)$ and described as the right process time chart in Figure 2.

F-included Before (fb)/After (fa). bf-relations *f-included before/after* are represented by bf-annotations fb$(5, 7)$/fa $(7, 5)$ and described as the left process time chart in Figure 4.

Paraconsistent Before After (pba). bf-relations *paraconsistent before after* are represented by bf-annotations pba $(6, 6)$ and described as the right process time chart in Figure 4.

If we take before-after measure over the ten bf-annotations as the horizontal order and before-after knowledge amount of them as the vertical one, we obtain the complete bi-lattice $\mathcal{T}_v(12)_{bf}$ of bf-annotations in Figure 5.

3 Reasoning Systems for Before-After Relations

First of all, we introduce an example for reasoning bf-relations between processes Pr_0, Pr_1 and Pr_2 in the process schedule chart in Figure 5, and describe how the bf-relations are infered at each time $t_0, \cdots, t_7$. Hereafter, we omit deontic annotations in bf-EVALPSN for simplicity.

At time t_0, no process has started, thus we have no knowledge about bf-relations between processes Pr_0, Pr_1 and Pr_2. Therefore, we have the bf-EVALPSN clauses,

$$R(Pr_0, Pr_1, t_0) : (0, 0), \quad R(Pr_1, Pr_2, t_0) : (0, 0), \quad R(Pr_0, Pr_2, t_0) : (0, 0).$$

At time t_1, only process Pr_0 has started, then it can be reasoned that the bf-relation between processes Pr_0 and Pr_1 is one of bf-relations $\{$db$(0, 12)$, mb$(1, 11)$, jb$(2, 10)$, sb$(3, 9)$, ib$(4, 8)\}$ whose greatest lower bound is $(0, 8)$. On the other hand we still have no knowledge about the bf-relation between processes Pr_1 and Pr_2. Moreover, literal $R(Pr_0, Pr_2, t_1)$ has the same vector annotation $(0, 8)$ as literal $R(Pr_0, Pr_1, t_1)$, and we have the bf-EVALPSN clauses,

$$R(Pr_0, Pr_1, t_0) : (0, 8), \quad R(Pr_1, Pr_2, t_0) : (0, 0), \quad R(Pr_0, Pr_2, t_0) : (0, 8).$$

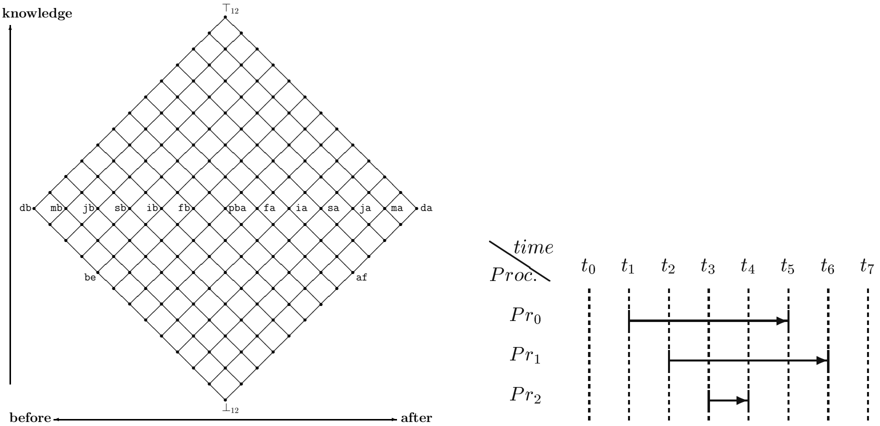


Fig. 5. Complete Lattice $\mathcal{T}_v(12)_{bf}$ and Process Schedule Chart

At time t_2, process Pr_1 has started before process Pr_0 finishes, then it can be reasoned that the bf-relation between processes Pr_0 and Pr_1 is one of bf-relations $\{\texttt{jb}(2,10), \texttt{sb}(3,9), \texttt{ib}(4,8)\}$ whose greatest lower bound is $(2,8)$. Moreover, as process Pr_2 has not started yet, we have the following bf-EVALPSN clauses,

$$R(Pr_0, Pr_1, t_0):(2,8), \quad R(Pr_1, Pr_2, t_0):(0,8), \quad R(Pr_0, Pr_2, t_0):(0,8).$$

At time t_3, process Pr_2 has started before processes Pr_0 and Pr_1 finish, then it can be reasoned that all literals $R(Pr_0, Pr_1, t_3)$, $R(Pr_1, Pr_2, t_3)$ $R(Pr_0, Pr_2, t_3)$ have the same vector annotation $(2,8)$ as well as literal $R(Pr_0, Pr_1, t_2)$. Therefore, we have the bf-EVALPSN clauses,

$$R(Pr_0, Pr_1, t_0):(2,8), \quad R(Pr_1, Pr_2, t_0):(2,8), \quad R(Pr_0, Pr_2, t_0):(2,8).$$

At time t_4, only process Pr_2 has finished, then only literal $R(Pr_0, Pr_1, t_4)$ still has the same vector annotation $(2,8)$, and it can be reasoned that literals $R(Pr_1, Pr_2,\ t_4)$ and $R(Pr_0, Pr_2, t_4)$ has their final vector annotation $\texttt{ib}(4,8)$. Therefore, we have the bf-EVALPSN clauses,

$$R(Pr_0, Pr_1, t_0):(2,8), \quad R(Pr_1, Pr_2, t_0):(4,8), \quad R(Pr_0, Pr_2, t_0):(4,8).$$

At time t_5, process Pr_0 has finished before processes Pr_1 finish, then literal $R(Pr_1, Pr_2, t_5)$ has its final vector annotation $\texttt{jb}(2,10)$. Therefore, even though process Pr_1 has not finished yet, all bf-relations between processes Pr_0, Pr_1 and Pr_2 have been determined as follows:

$$R(Pr_0, Pr_1, t_0):\texttt{jb}(2,10), \quad R(Pr_1, Pr_2, t_0):\texttt{ib}(4,8), \quad R(Pr_0, Pr_2, t_0):\texttt{ib}(4,8).$$

3.1 Inference Rules for Bf-Relations between Two Processes

As shown in the example, it is quite natural to construct inference rules for reasoning bf-relations between two processes Pr_i and Pr_j in real time. We introduce two kinds of predicates st and fi representing start/finish information of processes as follows: $st(p_i, t)$, "process Pr_i starts at time t"; $fi(p_i, t)$, "process Pr_i finishes at time t". Those literals have vector annotations $\mathtt{t}(1, 0)/\mathtt{f}(0, 1)$ that mean true/false intuitively. We introduce the following inference rules.

BF-RELATION INFERENCE RULES

(0, 0)-rule-1 If process Pr_i has just started and process Pr_j has not started yet, the vector annotation of $R(p_i, p_j, t)$ becomes $(0, 8)$ from $(0, 0)$.

(0, 0)-rule-2 If both processes Pr_i and Pr_j have started at the same time, the vector annotation of $R(p_i, p_j, t)$ becomes $(5, 5)$ from $(0, 0)$.

These $(0, 0)$-rule-1,2 may be translated into the bf-EVALPSN clauses:

$$R(p_i, p_j, t) : (0, 0) \wedge st(p_i, t) : \mathtt{t} \wedge \sim st(p_j, t) : \mathtt{t} \rightarrow R(p_i, p_j, t) : (0, 8),$$
$$R(p_i, p_j, t) : (0, 0) \wedge st(p_i, t) : \mathtt{t} \wedge st(p_j, t) : \mathtt{t} \rightarrow R(p_i, p_j, t) : (5, 5).$$

(0,8)-rule-1 If process Pr_i has finished, and process Pr_j has not started yet, then the vector annotation of $R(p_i, p_j, t)$ becomes $(0, 12)$(disjoint before) from $(0, 8)$.

(0,8)-rule-2 If process Pr_i has finished, and process Pr_j has started immediately after process Pr_i finishing, then the vector annotation of $R(p_i, p_j, t)$ becomes $(1, 11)$(immediate before) from $(0, 8)$.

(0,8)-rule-3 If process Pr_i has started but not finished yet, and process Pr_j has also started after process Pr_i starting, then the vector annotation of $R(p_i, p_j, t)$ becomes $(2, 8)$ from $(0, 8)$.

These $(0, 8)$-rule-1,2,3 may be translated into the bf-EVALPSN clauses:

$$R(p_i, p_j, t) : (0, 8) \wedge fi(p_i, t) : \mathtt{t} \wedge \sim st(p_j, t) : \mathtt{t} \rightarrow R(p_i, p_j, t) : (0, 12),$$
$$R(p_i, p_j, t) : (0, 8) \wedge fi(p_i, t) : \mathtt{t} \wedge st(p_j, t) : \mathtt{t} \rightarrow R(p_i, p_j, t) : (1, 11),$$
$$R(p_i, p_j, t) : (0, 8) \wedge \sim fi(p_i, t) : \mathtt{t} \wedge st(p_j, t) : \mathtt{t} \rightarrow R(p_i, p_j, t) : (2, 8).$$

(5,5)-rule-1 If both processes Pr_i and Pr_j have started simultaneously and only process Pr_i has finished, then the vector annotation of $R(p_i, p_j, t)$ becomes $(5, 7)$(s-included before) from $(5, 5)$.

(5,5)-rule-2 If both processes Pr_i and Pr_j have started simultaneously and finished simultaneously, then the vector annotation of $R(p_i, p_j, t)$ becomes $(6, 6)$(paraconsistent before/after) from $(5, 5)$.

(5,5)-rule-3 If both processes Pr_i and Pr_j have started simultaneously and only process Pr_j has finished, then the vector annotation of $R(p_i, p_j, t)$ becomes $(7, 5)$(s-included after) from $(5, 5)$.

These $(5,5)$-rule-1,2,3 may be translated into the bf-EVALPSN clauses:

$$R(p_i,p_j,t):(5,5) \wedge fi(p_i,t):\mathtt{t} \wedge \sim fi(p_j,t):\mathtt{t} \rightarrow R(p_i,p_j,t):(5,7),$$
$$R(p_i,p_j,t):(5,5) \wedge fi(p_i,t):\mathtt{t} \wedge fi(p_j,t):\mathtt{t} \rightarrow R(p_i,p_j,t):(6,6),$$
$$R(p_i,p_j,t):(5,5) \wedge \sim fi(p_i,t):\mathtt{t} \wedge fi(p_j,t):\mathtt{t} \rightarrow R(p_i,p_j,t):(7,5).$$

(2,8)-rule-1 If processes Pr_i and Pr_j have started sequentially, process Pr_i has finished, and process Pr_j has not finished yet, then the vector annotation of $R(p_i,p_j,t)$ becomes $(2,10)$(joint before) from $(2,8)$.

(2,8)-rule-2 If processes Pr_i and Pr_j have started sequentially and they finished at the same time, then the vector annotation of $R(p_i,p_j,t)$ becomes $(3,9)$(f-included before) from $(2,8)$.

(2,8)-rule-3 If processes Pr_i and Pr_j have started sequentially and process Pr_i has not finished yet, though process Pr_j has finished yet, then the vector annotation of $R(p_i,p_j,t)$ becomes $(4,8)$(included before) from $(2,8)$.

These $(2,8)$-rule-1,2,3 may be translated into the bf-EVALPSN clauses:

$$R(p_i,p_j,t):(2,8) \wedge fi(p_i,t):\mathtt{t} \wedge \sim fi(p_j,t):\mathtt{t} \rightarrow R(p_i,p_j,t):(2,10),$$
$$R(p_i,p_j,t):(2,8) \wedge fi(p_i,t):\mathtt{t} \wedge fi(p_j,t):\mathtt{t} \rightarrow R(p_i,p_j,t):(3,9),$$
$$R(p_i,p_j,t):(2,8) \wedge \sim fi(p_i,t):\mathtt{t} \wedge fi(p_j,t):\mathtt{t} \rightarrow R(p_i,p_j,t):(4,8).$$

3.2 Transitive Reasoning System of Bf-Relations

We suppose three processes Pr_i,Pr_j and Pr_k starting sequentially. Then, we consider some real-time inference rules to reason the vector annotation of $R(p_i,p_k,t)$ from vector annotation of $R(p_i,p_j,t$ and $R(p_j,p_k,t)$ based on the complete lattice $\mathcal{T}_v(12)_{bf}$ in Figure 5.

We introduce how to construct the inference rules with taking three examples.

Case 1. Suppose that only process Pr_i has started at time t, we obtain the vector annotation $(0,8)$ of $R(p_i,p_j,t)$ by the $(0,0)$-rule-1 and the vector annotation $(0,0)$ of $R(p_j,p_k,t)$, then the vector annotation of $R(p_i,p_k,t)$ can be reasoned deterministically as $(0,8)$, therefore we have the bf-EVALP clause rule,

$$R(p_i,p_j,t):(0,8) \wedge R(p_j,p_k,t):(0,0) \rightarrow R(p_i,p_j,t):(0,8).$$

Case 2. Suppose that processes Pr_i and Pr_j have started simultaneously at time t, we obtain the vector annotation $(5,5)$ of $R(p_i,p_j,t)$ by the $(0,0)$-rule-2 and the vector annotation $(0,8)$ of $R(p_j,p_k,t)$ by the $(0,0$-rule-1, then the vector annotation of $R(p_i,p_k,t)$ can be also reasoned deterministically as $(0,8)$, therefore we have the bf-EVALP clause rule,

$$R(p_i,p_j,t):(5,5) \wedge R(p_j,p_k,t):(0,8) \rightarrow R(p_i,p_j,t):(0,8).$$

Case 3. Suppose that all processes Pr_i, Pr_j and Pr_k have started simultaneously at time t, we obtain the same vector annotation $(5,5)$ of $R(p_i,p_j,t)$ and $R(p_j,p_k,t)$ by the $(0,0)$-rule-2, then the vector annotation of $R(p_i,p_k,t)$

can be also reasoned deterministically as $(5,5)$, therefore we have the bf-EVALP clause rule,

$$R(p_i,p_j,t):(5,5) \wedge R(p_j,p_k,t):(5,5) \rightarrow R(p_i,p_j,t):(5,5).$$

Now we show all transitive inference rules for bf-relations in bf-EVALP systematically. For simplicity we represent the inference rule by thier vector annotations as follows: $(n_1,n_2) \wedge (n_3,n_4) \rightarrow (n_5,n_6)$, instead of the full bf-EVALP clause rule, $R(p_i,p_j,t):(n_1,n_2) \wedge R(p_j,p_k,t):(n_3,n_4) \rightarrow R(p_i,p_k,t):(n_5,n_6)$.

TRANSITIVE INFERENCE RULES

$\mathbf{R1}$ $(0,8) \wedge (0,0) \rightarrow (0,8)$

$\mathbf{R1-1}$ $(0,12) \wedge (0,0) \rightarrow (0,12)$

$\mathbf{R1-2}$ $(2,8) \wedge (0,8) \rightarrow (0,8)$

$\mathbf{R1-2-1}$ $(2,10) \wedge (0,8) \rightarrow (0,12)$

$\mathbf{R1-2-2}$ $(4,8) \wedge (0,12) \rightarrow (0,8)$ (1)

$\mathbf{R1-2-3}$ $(2,8) \wedge (2,8) \rightarrow (2,8)$

$\mathbf{R1-2-3-1}$ $(2,10) \wedge (2,8) \rightarrow (2,10)$

$\mathbf{R1-2-3-2}$ $(4,8) \wedge (2,10) \rightarrow (2,8)$ (2)

$\mathbf{R1-2-3-3}$ $(2,8) \wedge (4,8) \rightarrow (4,8)$

$\mathbf{R1-2-3-4}$ $(3,9) \wedge (2,10) \rightarrow (2,10)$

$\mathbf{R1-2-3-5}$ $(2,10) \wedge (4,8) \rightarrow (3,9)$

$\mathbf{R1-2-3-6}$ $(4,8) \wedge (3,9) \rightarrow (4,8)$

$\mathbf{R1-2-3-7}$ $(3,9) \wedge (3,9) \rightarrow (3,9)$

$\mathbf{R1-2-4}$ $(3,9) \wedge (0,12) \rightarrow (0,12)$

$\mathbf{R1-2-5}$ $(2,10) \wedge (2,8) \rightarrow (1,11)$

$\mathbf{R1-2-6}$ $(4,8) \wedge (1,11) \rightarrow (2,8)$ (3)

$\mathbf{R1-2-7}$ $(3,9) \wedge (1,11) \rightarrow (1,11)$

$\mathbf{R1-3}$ $(1,11) \wedge (0,8) \rightarrow (0,12)$

$\mathbf{R1-4}$ $(2,8) \wedge (5,5) \rightarrow (2,8)$

$\mathbf{R1-4-1}$ $(4,8) \wedge (5,7) \rightarrow (2,8)$ (4)

$\mathbf{R1-4-2}$ $(2,8) \wedge (7,5) \rightarrow (4,8)$

$\mathbf{R1-4-3}$ $(3,9) \wedge (5,7) \rightarrow (2,10)$

$\mathbf{R1-4-4}$ $(2,10) \wedge (7,5) \rightarrow (3,9)$

$\mathbf{R2}$ $(5,5) \wedge (0,8) \rightarrow (0,8)$

$\mathbf{R2-1}$ $(5,7) \wedge (0,8) \rightarrow (0,12)$

$\mathbf{R2-2}$ $(7,5) \wedge (0,12) \rightarrow (0,8)$ (5)

$\mathbf{R2-3}$ $(5,5) \wedge (2,8) \rightarrow (2,8)$

$\mathbf{R2-3-1}$ $(5,7) \wedge (2,8) \rightarrow (2,10)$

$$\mathbf{R2-3-2} \quad (7,5) \wedge (2,10) \rightarrow (2,8) \tag{6}$$

$$\mathbf{R2-3-3} \quad (5,5) \wedge (4,8) \rightarrow (4,8)$$

$$\mathbf{R2-3-4} \quad (7,5) \wedge (3,9) \rightarrow (4,8)$$

$$\mathbf{R2-4} \quad (5,7) \wedge (2,8) \rightarrow (1,11)$$

$$\mathbf{R2-5} \quad (7,5) \wedge (1,11) \rightarrow (2,8) \tag{7}$$

$$\mathbf{R3} \quad (5,5) \wedge (5,5) \rightarrow (5,5)$$

$$\mathbf{R3-1} \quad (7,5) \wedge (5,7) \rightarrow (5,5) \tag{8}$$

$$\mathbf{R3-2} \quad (5,7) \wedge (7,5) \rightarrow (6,6)$$

The above inference rule names indicate applicable orders of the rules. For example, if rule **R1** has been applied, the next applicable rules are rules **R1-1**, **R1-2**, **R1-3**, and **R1-4**. Furthermore, if rules **R1-2** or **R1-4** are applied, rules **R1-2-1**,$\cdots$,**R1-2-7**, or **R1-4-1**,$\cdots$,**R1-4-4** can be applied; on the other hands, if rules **R1-1** or **R1-3** are applied, the bf-relation between processes Pr_i and Pr_k are determined. Forcusing on the inference rules (1), (2), (3), (4), (5), (6), (7) and (8), even though they have no following rules to be applied, they can not derive the bf-annotation of $R(p_i, p_k, t)$. For example, in the case of rule **R-1-2-3-2**, even if both the bf-annotations $(4,8)$ of processes Pr_i and Pr_j, and $(2,10)$ of processes Pr_j and Pr_k are obtained, the bf-annotation of Pr_i and Pr_k is still undetermined and it is just $(2,8)$ that has the possibilities to be one of $\{(2,10),(3,9),(4,8)\}$. Therefore, if those inference rules have been applied, rules (0,8),(2,8),(5,5)-rules that can reason bf-relations between two processes introduced previously have to be applied for reasoning the bf-annotation of processes Pr_i and Pr_k. For example, if rule **R-1-2-3-2** has been applied, (2,8)-rules should be applied.[2]

References

1. Nakamatsu, K., Abe, J.M., Suzuki, A.: Annotated Semantics for Defeasible Deontic Reasoning. In: Ziarko, W.P., Yao, Y. (eds.) RSCTC 2000. LNCS (LNAI), vol. 2005, pp. 432–440. Springer, Heidelberg (2001)
2. Nakamatsu, K.: Pipeline Valve Control Based on EVALPSN Safety Verification. J. Advanced Computational Intelligence and Intelligent Informatics 10, 647–656 (2006)
3. Nakamatsu, K., Mita, Y., Shibata, T.: An Intelligent Action Control System Based on Extended Vector Annotated Logic Program and its Hardware Implementation. J. Intelligent Automation and Soft Computing 13, 289–304 (2007)
4. Nakamatsu, K., Abe, J.M., Akama, S.: Paraconsistent Before-after Relation Reasoning Based on EVALPSN. In: New Directions in Intelligent Interactive Multimedia. Studies in Computational Intelligence, vol. 142, pp. 265–274. Springer, Heidelberg (2008)
5. Nakamatsu, K., Akama, S., Abe, J.M.: Annotated Semantics for Defeasible Deontic Reasoning. In: Lovrek, I., Howlett, R.J., Jain, L.C. (eds.) KES 2008, Part II. LNCS (LNAI), vol. 5178, pp. 474–482. Springer, Heidelberg (2008)
6. Nakamatsu, K., Abe, J.M.: The development of Paraconsistent Annotated Logic Program. Int'l J. Reasoning-based Intelligent Systems, vol. 1 (to appear, 2009)

[2] This work is financially supported by Japanese Scientific Research Grant (C) No. 20500074.

A Note on Monadic Curry System P_1

Jair Minoro Abe[1,2], Kazumi Nakamatsu[3], and Fábio Romeu de Carvalho[1]

[1] Graduate Program in Production Engineering, ICET - Paulista University
R. Dr. Bacelar, 1212, CEP 04026-002 São Paulo – SP – Brazil
fabioromeu@unip.br
[2] Institute For Advanced Studies – University of São Paulo, Brazil
jairabe@uol.com.br
[3] School of Human Science and Environment/H.S.E. – University of Hyogo – Japan
nakamatu@shse.u-hyogo.ac.jp

Abstract. This paper is a sequel to [4]. We present an algebraic version of the monadic system P_1* [8] by using the concept of Curry Algebra [5]. The algebraic structure obtained is called Monadic Curry Algebra P_1*, which is a kind of 'dual' algebra studied in [4].

Keywords: Curry algebras, system P_1, paracomplete logic, algebraic logic, non-classical logic.

1 Introduction

The appearance of some non-classical logics constitutes a landmark in the History of Logic over the past decades. One of them, the so-called paraconsistent logic, is a logic that, roughly speaking, allows contradictions without trivialization in theories based on it. So, in paraconsistent logics, the principle of the contradiction (from two contradictory formulas A and $\neg A$ (the negation of A), one must be false) fails. One of the first important paraconsistent systems studied in the literature is the system C_n ($1 \leq n < \omega$) [7]. One natural question is the algebraic version of these systems. Several works were presented, including using the concept of Curry system [5]. The 'dual' system of paraconsistent logics are known as paracomplete logics. So, in these systems, the principle of the excluded middle (from two contradictory formulas A and $\neg A$, one must be true) fails. In [8] it was introduced a hierarchy of paracomplete systems P_n ($1 \leq n < \omega$), dual in a precise sense of the systems C_n ($1 \leq n < \omega$).

The purpose of this paper is to study an algebraic version of the paracomplete system P_1 via the concept of Curry algebra.

2 Background

We begin with basic concepts. For a detailed account see [5].

Definition 1. A system $<A, \equiv, \leq>$ is called a pre-ordered system if for all $x \in A$, $x \leq x$; for all $x, y, z \in A$, $x \leq y$ and $y \leq z$ imply $x \leq z$; for all $x, y, x', y' \in A$, $x \leq y$, $x \equiv x'$, and $y \equiv y'$ imply $x' \leq y'$.

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 388–394, 2009.

A pre-ordered system $<A, \equiv, \leq>$ is called a partially-ordered system if for all $x, y \in A$, $x \leq y$ and $y \leq x$ imply $x \equiv y$;

A partially-ordered system $<A, \equiv, \leq>$ is called a pre-lattice system if for all $x, y \in A$, the set of $\sup\{x, y\} \neq \varnothing$ and the set of $\inf\{x, y\} \neq \varnothing$. We denote by $x \vee y$ one element of the set of $\sup\{x, y\}$ and by $x \wedge y$ one element of the set of $\inf\{x, y\}$.

A system $<A, \equiv, \leq, \rightarrow>$ is called an implicative pre-lattice if $<A, \equiv, \leq>$ is a pre-lattice, and for all $x, y, z \in A$, $x \wedge (x \rightarrow y) \leq y$ and $x \wedge y \leq z$ iff $x \leq y \rightarrow z$.

$<A, \equiv, \leq, \rightarrow>$ is called classical implicative pre-lattice if it is an implicative pre-lattice and $(x \rightarrow y) \rightarrow x \leq x$ (Peirce's law).

With this definition we can extend the majority of algebraic systems to pre-algebraic systems considering an equivalence relation $\equiv$ instead of equality relation. In this way we can obtain, v.g., the concepts of Boolean pre-algebras, pre-filters, pre-lattices, etc.

Let's consider the logical systems P_n $(1 \leq n < \omega)$ [8].

The primitive symbols of the language L of the calculi P_1 are the following:

- Propositional variables: a denumerable set of propositional variables;
- Logical connectives: $\rightarrow$ (implication), $\wedge$ (conjunction), $\vee$ (disjunction), $\neg$ (negation);
- Parentheses.

Formulas are defined in the usual manner. In L, we put:

Definition 2. Let A be any formula. Then $A^{\#}$ is shorthand for $A \vee \neg A$. Also, we write $A \leftrightarrow B$ for $(A \rightarrow B) \wedge (B \rightarrow A)$.

The postulates (axiom schemes and inference rules) of P_1 are: A, B, and C are formulas whatsoever.

(1) $A \rightarrow (B \rightarrow A)$

(2) $(A \rightarrow (B \rightarrow C)) \rightarrow ((A \rightarrow B) \rightarrow (A \rightarrow C))$

(3) $\dfrac{A,\ A \rightarrow B}{B}$

(4) $A \wedge B \rightarrow A$

(5) $A \wedge B \rightarrow B$

(6) $A \rightarrow (B \rightarrow (A \wedge B))$

(7) $A \rightarrow A \vee B$

(8) $B \rightarrow A \vee B$

(9) $(A \rightarrow C) \rightarrow ((B \rightarrow C) \rightarrow ((A \vee B) \rightarrow C))$

(10) $A^{\#} \rightarrow ((A \rightarrow B) \rightarrow ((A \rightarrow \neg B) \rightarrow \neg A)$

(11) $A^{\#} \wedge B^{\#} \rightarrow (A \wedge B)^{\#} \wedge (A \vee B)^{\#} \wedge (A \rightarrow B)^{\#} \wedge (\neg A)^{\#}$

(12) $\neg(\neg A \wedge A)$

(13) $A \rightarrow (\neg A \rightarrow B)$

(14) $A \rightarrow \neg\neg A$

In P_1, $A^{\#}$ expresses intuitively that the formula A 'behaves' classically, so that the motivation of the postulates (10) and (11) are clear. Furthermore, in this calculus, the set of all well-behaved formulas together with the connectives $\rightarrow$, $\wedge$, $\vee$, and $\neg$ have all the properties of classical implication, conjunction, disjunction, and negation, respectively. Therefore the classical propositional calculus is contained in P_1, though it constitutes a strict sub-calculus of the former.

Theorem 1. In P_1, all valid schemes and rules of classical positive propositional logic are true. In particular, the deduction theorem is valid in P_1.

Theorem 2. In P_1, the following schemes are not valid (among others): $A \vee \neg A$; $\neg\neg A \leftrightarrow A$; $\neg(A \vee B) \leftrightarrow \neg A \wedge \neg B$; $\neg(A \wedge B) \leftrightarrow \neg A \vee \neg B$; $(A \to B) \to (\neg B \to \neg A)$; $A^{\#\#}$.

Theorem 3. In P_1 we have: $\vdash (A \wedge \neg A) \to B$; $\vdash A \vee (A \to B)$, $\vdash A^{\#} \to (\neg\neg A \leftrightarrow A)$; $\vdash A^{\#} \to ((A \to B) \to (\neg A))$; $\vdash (A \wedge \neg A) \leftrightarrow (B \wedge \neg B)$.

Let's define the operator $\neg^{*}A =_{\text{Def.}} A \to (B \wedge \neg B)$ (where B is a fixed formula). Such operator $\neg^{*}$ is called strong negation and with the remaining connectives $\wedge$, $\vee$, and $\to$ they have all properties of classical connectives of negation, conjunction, disjunction, and implication, respectively. In short, the classical logic is contained in P_1.

3 The Curry Algebra P_1

Definition 3. A Curry algebra P_1 (or a P_1-algebra) is a classical implicative pre-lattice $<A, \equiv, \leq, \wedge, \vee, \to, '>$ with a greatest element 1 and operators $\wedge$, $\vee$, and ' satisfying the conditions below, where $x^{\#} =_{\text{Def.}} x \vee x'$:

(1) $x \leq x''$

(2) $(x \wedge x') \equiv 1$

(3) $x^{\#} \wedge y^{\#} \leq (x \to y)^{\#}$;

(4) $x^{\#} \wedge y^{\#} \leq (x \wedge y)^{\#}$;

(5) $x^{\#} \wedge y^{\#} \leq (x \vee y)^{\#}$;

(6) $x^{\#} \leq (x')^{\#}$

(7) $x^{\#} \leq (x \to y) \to ((x \to y') \to x')$

(8) $x \leq (x' \to y)$

Example 1. Let's consider the calculus P_1. A is the set of all formulas of P_1. Let's consider as operations, the logical connectives of conjunction, disjunction, implication, and negation. Let's define the relation on A:

$x \equiv y$ iff $\vdash x \leftrightarrow y$. It is easy to check that $\equiv$ is an equivalence relation on A.

$x \leq y$ iff $x \equiv x \wedge y$ and $y \leq x$ iff $y \equiv x \wedge y$. Also we take as 1 any fixed axiom instance.

The structure composed $<A, \equiv, \leq, \wedge, \vee, \to, '>$ is a P_1-algebra.

Theorem 4. Let's $<A, \equiv, \leq, \wedge, \vee, \to, '>$ be a P_1-algebra. Then the operator ' is non-monotone relatively $\equiv$.

Theorem 5. A P_1-algebra is distributive and has a greatest element, as well as a first element.

Definition 4. Let x be an element of a P_1-algebra. We put $x^* = x \to (y \wedge y')$, where y is a fixed element.

Theorem 6. In a P_1-algebra, x^* is a Boolean complement of x; so $x \vee x^* \equiv 1$ and $x \wedge x^* \equiv 0$. Moreover, in a P_1-algebra, the structure composed by the underlying set and by operations $\wedge$, $\vee$, and * is a (pre) Boolean algebra. If we pass to the quotient by the basic relation $\equiv$, we obtain a Boolean algebra in the usual sense.

Definition 5. Let $<A, \equiv, \leq, \wedge, \vee, \rightarrow, '>$ be a P_1-algebra and $<A, \equiv, \leq, \wedge, \vee, \rightarrow, *>$ the Boolean algebra obtained as in the above theorem. Any Boolean algebra that is isomorphic to the quotient algebra of $<A, \equiv, \leq, \wedge, \vee, \rightarrow, *>$ by $\equiv$ is called Boolean algebra *associated with the* P_1-algebra.

Hence, we have the following representation theorems for P_1-algebras.

Theorem 7. Any P_1-algebra is associated with a field of sets. Moreover, any P_1-algebra is associated with the field of sets simultaneously open and closed of a totally disconnected compact Hausdorff space.

One open problem concerning P_1-algebras remains. How many non-isomorphic associated with the P_1-algebra are there?

4 The P_n-Algebras

Now we show a chain of Curry algebras beginning with the P_1-algebra.

Let $<A, \equiv, \leq, \wedge, \vee, \rightarrow, '>$ be a P_1-algebra. If $x \in A$, x^1 abbreviates $x^\#$. x^n ($1 < n < \omega$) abbreviates $x^\# \wedge x^{\#\#} \wedge ... \wedge x^{\#\#...\#}$, where the symbol $^\#$ occurs n times. Also, $x^{(n)}$ abbreviates $x^1 \wedge x^2 \wedge ... \wedge x^n$.

Definition 6. A P_n-algebra ($1 < n < \omega$) is an implicative pre-lattice $<A, \equiv, \leq, \wedge, \vee, \rightarrow, '>$ with a first element 1 and operators $\wedge$, $\vee$, and ' satisfying the conditions:

(1) $x \leq x''$

(2) $(x \wedge x') \equiv 1$

(3) $x^{(n)} \wedge y^{(n)} \leq (x \rightarrow y)^{(n)}$;

(4) $x^{(n)} \wedge y^{(n)} \leq (x \wedge y)^{(n)}$;

(5) $x^{(n)} \wedge y^{(n)} \leq (x \vee y)^{(n)}$;

(6) $x^{(n)} \leq (x')^{(n)}$

(7) $x^{(n)} \leq (x \rightarrow y) \rightarrow ((x \rightarrow y') \rightarrow x')$

(8) $x \leq (x' \rightarrow y)$

Usual algebraic structural concepts like homomorphism, monomorphism, etc. can be introduced for Curry algebras without extensive comments.

Theorem 8. Every P_n-algebra is embedded in any P_{n-1}-algebra ($1 < n < \omega$).

Corollary 8.1. Every P_n-algebra ($1 < n < \omega$) is embedded in any P_1-algebra.

If we indicate a P_n-algebra by P_n, the embedding hierarchy can be represented as $P_1 > P_2 > \dots P_n > \dots$

Definition 7. A P_ω-algebra is an implicative pre-lattice $<A, \equiv, \leq, \wedge, \vee, \rightarrow, '>$ with a first element 1 and operators $\wedge$, $\vee$, and ' satisfying the conditions below:

(1) $(x \wedge x')' \equiv 1$

(2) $x \leq (x' \rightarrow y)$

(3) $x \leq x''$

We propose in the sequence some extensions of the P_1-algebras.

5 P_1*-Monadic Algebras

In this section we present some monadic Curry algebras P_1*.

Definition 8. Let A be a P_1-algebra. Let $\exists$ and $\forall$ be operators on A. ($\exists$, $\forall$) is called a quantifier on A if

(1) $\exists 0 \equiv 0$;
(2) $x \leq \exists x$;
(3) $\exists(x \vee y) \equiv \exists x \vee \exists y$;
(4) $\exists\exists x \equiv \exists x$;
(5) $\exists(\exists x)^* \equiv (\exists x)^*$;
(6) $\exists(x \wedge \exists y) \equiv \exists x \wedge \exists y$;
(7) $\forall 1 \equiv 1$;
(8) $\forall x \leq x$;
(9) $\forall(x \vee y) \equiv \forall x \vee \forall y$;
(10) $\forall\forall x \equiv \forall x$;
(11) $\forall(\forall x)^* \equiv (\forall x)^*$

We suppose in the above definition that, if $x \equiv y$, then $\exists x \equiv \exists y$ and $\forall x \equiv \forall y$. $\exists$ is called existential quantifier on A and $\forall$ is called universal quantifier on A. The pair <A, ($\exists$, $\forall$)> is called a monadic Curry algebra P_1* or a P_1*-monadic algebra (or P_1*-algebra).

Given a Curry algebra P_1, let's assume that there is an universal quantifier defined on it, i.e., a structure <A, $\forall$> such that conditions (7) - (11) above are satisfied. If we define $\exists_1 x =_{\text{Def.}} (\forall x^*)^*$, then $\exists_1$ is an existential quantifier (i.e. satisfying (1) – (6) and the structure composed by <A, ($\exists_1$, $\forall$)> is a P_1*-monadic algebra. Also, we can get a monadic algebra considering an existential quantifier $\exists$ on a Curry algebra P_1 satisfying conditions (1) – (6) of above definition and defining a universal quantifier (i.e., satisfying (7) – (11) as $\forall_1 x =_{\text{Def.}} (\exists x^*)^*$. Then the structure composed by <A, ($\exists$, $\forall_1$)> is a P_1*-monadic algebra. Given a Curry algebra P_1, in general, the algebras obtained <A, ($\exists_1$, $\forall$)> and <A, ($\exists$, $\forall_1$)> are not isomorphic. Also, given a P_1*-monadic algebra <A, ($\exists$, $\forall$)>, define the new quantifiers $\exists_1$ and $\forall_1$. In general, we have $\exists_1 \neq \exists$ and $\forall_1 \neq \forall$.

Let P be a P_1-algebra and $A = \{(x_1, x_2, \dots, x_n) | x_i \in C, i = 1, 2, \dots, n\}$. Let us suppose that if $x, y \in A$, then $x \vee y \in A$ and $x' \in A$, and if $(x_1, x_2, \dots, x_n), (y_1, y_2, \dots, y_n) \in A$ we define $(x_1, x_2, \dots, x_n) \equiv (y_1, y_2, \dots, y_n)$ iff $x_i \equiv y_i$, $i = 1, 2, \dots, n$. Also, we put $\overline{x} = \bigvee_{i=1}^{n} x_i$, $\underline{x} = \bigwedge_{i=1}^{n} x_i$, and we assume that $\underbrace{(\overline{x}, \dots, \overline{x})}_{n-times}$, $\underbrace{(\underline{x}, \dots, \underline{x})}_{n-times} \in A$. If we define $\exists x = (\overline{x}, \dots, \overline{x})$ and $\forall x = (\underline{x}, \dots, \underline{x})$, then A is a C_1*-monadic algebra.

A more useful example of P_1*-monadic algebra is the following. Let P be a P_1-algebra, a set $K \neq \varnothing$, and P^K is the set of all functions of K into P. Let A be the set such that: (i) A is a P_1-algebra with respect the pointwise operations, and (ii) if $x \in A$, then the range of x has a supremum $\overline{x}$ and a infimum $\underline{x}$ in P, and the functions that take the value $\overline{x}$ at each point of K and $\underline{x}$ at each point of K are in A. If $\exists x$ and $\forall x$ are defined to be those functions, then A becomes a P_1*-monadic algebra. Every P_1*-monadic algebra obtained in this way is called a P-valued functional algebra with domain K.

Now we discuss some results of the algebraic structures originated in our discussion. In fact, the matter is very rich, but due limitations of this paper we'll concern only on some of them.

Theorem 9. In a P_1*-monadic algebra $<A, (\exists, \forall)>$, the structure composed by the underlying set and by operations $\wedge$, $\vee$, *, $\exists$, and $\forall$ is a (pre) monadic algebra. If we pass to the quotient by the basic relation $\equiv$, we obtain a monadic algebra in the usual sense [10].

Definition 9. Let $<A, (\exists, \forall)>$ be a P_1*-monadic algebra, and $<A, \equiv, \leq, \rightarrow, *, \exists, \forall>$ the monadic algebra obtained as in the above theorem. Any monadic algebra that is isomorphic to the quotient algebra of $<A, \equiv, \leq, \rightarrow, *, \exists, \forall>$ by $\equiv$ is called monadic algebra *associated with the* P_1*-monadic algebra.

Hence, we can establish the following representation theorems for P_1*-monadic algebras.

Theorem 10. If P is a P_1*-monadic algebra, then for its associated monadic algebra A, there exists a set X and there exists a Boolean algebra B, such that (i) A is isomorphic to a B-valued functional algebra A' with domain X, and (ii) for every element p of A' there exists a point x in X with $p(x) = \exists p(x)$.

Theorems 9 and 10 show us that P_1*-monadic algebras constitute interesting generalization of the concept of monadic algebras. Here, there is an open problem. How many non-isomorphic monadic algebras associated with a P_1*-monadic algebra are there?

6 Conclusions

In this work we've studied the P_1-algebras which constitute the dual algebras of the C_1-algebras. It constitutes a promising area of researching with many interesting results to coming. We hope to say more in forthcoming papers.

References

1. Abe, J.M.: Curry algebras Pt. Logique et Analyse, 161–163, 5–15 (1998)
2. Abe, J.M.: Curry Algebras N1. Atti Acc. Lincei Rend. Fis. 7(9), 125–128 (1996)
3. Abe, J.M., Akama, S., Nakamatsu, K.: Monadic Curry Algebras Qt. In: Apolloni, B., Howlett, R.J., Jain, L. (eds.) KES 2007, Part II. LNCS, vol. 4693, pp. 893–900. Springer, Heidelberg (2007)
4. Abe, J.M., Nakamatsu, K., Akama, S.: An Algebraic Version of the Monadic System C1. To appear in the First KES International Symposium on Intelligent Decision Technologies (IDT 2009), Japan (2009)

5. Barros, C.M., da Costa, N.C.A., Abe, J.M.: Tópico de teoria dos sistemas ordenados: vol. II, sistemas de Curry, Coleção Documentos, Série Lógica e Teoria da Ciência, IEA-USP, vol. 20, 132 p. (1995)
6. Curry, H.B.: Foundations of Mathematical Logic. Dover, New York (1977)
7. Da Costa, N.C.A.: On the theory of inconsistent formal systems. Notre Dame J. of Formal Logic 15, 497–510 (1974)
8. Da Costa, N.C.A., Marconi, D.: A note on paracomplete logic. Atti Acc. Lincei Rend. Fis. 80(8), 504–509 (1986)
9. Eytan, M.: Tableaux of Hintikka et Tout ça: un Point de Vue Algebrique. Math. Sci. Humaines 48, 21–27 (1975)
10. Halmos, P.R.: Algebraic Logic. Chelsea Publishing Co., New York (1962)
11. Kleene, S.C.: Introduction to Metamathematics. Van Nostrand, Princeton (1952)
12. Mortensen, C.: Every quotient algebra for C1 is trivial. Notre Dame J. of Formal Logic 21, 694–700 (1977)

Adaptation of Space-Mapping Methods for Object Location Estimation to Camera Setup Changes — A New Study*

Chih-Jen Wu[1] and Wen-Hsiang Tsai[1,2]

[1] Institute of Computer Science and Engineering, National Chiao Tung University, Taiwan
gis91813@cis.nctu.edu.tw
[2] Department of Information Communication, Asia University, Taiwan
whtsai@cis.nctu.edu.tw.

Abstract. A new space-mapping method for object location estimation which is adaptive to camera setup changes in various applications is proposed. The location of an object appearing in an image is estimated by mapping image coordinates of object points to corresponding real-world coordinates using a mapping table, which is constructed in two stages, with the first for establishing a basic table using bilinear interpolation and the second for adapting it to changes of camera heights and orientations. Analytic equations for such adaptation are derived based on image formation and camera geometry properties. Good experimental results are shown to prove the feasibility of the proposed method.

Keywords: object location estimation, space mapping, table adaptation.

1 Introduction

Video cameras are used in various applications, including automatic estimation of the location of an object in an indoor environment using an object image acquired by a camera affixed to a wall or a ceiling. A conventional solution to this problem is to conduct camera calibration to obtain a set of camera parameters, followed by the use the parameters to compute the object location [1-5]. Camera calibration methods often use landmarks to compute camera parameters. The process is generally complicated. An alternative way is to use a *space-mapping table* [6-9] which transforms the image space into the real-world space, thus avoiding camera calibration. The table is constructed with the aid of a *calibration pattern* before the camera is deployed in an application environment.

Space-mapping based object location estimation however is *sensitive to camera setup changes*. That is, after a space-mapping table is constructed for a specific camera setup according to a certain camera-environment configuration, the camera should be used *in identical configurations thereafter*; otherwise, the table will not work. This weakness causes inconvenience in using the camera.

* This work was supported financially by the Ministry of Economic Affairs under Project No. MOEA 97-EC-17-A-02-S1-032 in Technology Development Program for Academia.

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 395–402, 2009.

To solve such a problem, one way is to construct a new table in a new camera-environment configuration. But this is often difficult to carry out after the camera is delivered to a user who does not know the table construction process. In this study, we investigate the possibility of automatically modifying the original space-mapping table for use in new environments. This is a new topic which has not been studied yet.

In the following, we first describe the idea and the detail of the proposed method in Sections 2 and 3, respectively. Some experimental results are given in Section 4, followed by conclusions in Section 5.

2 Idea of Proposed Method

The proposed method includes two stages, one conducted in an in-factory environment and the other in an in-field one. The details are described in the following algorithm. See Fig. 1 for an illustration.

Algorithm 1. Object location estimation by space-mapping table construction and modification.

Stage 1. Construction of a basic mapping table in the factory (see Fig. 1(a)).

Step. 1 Affix the camera to the ceiling at a height H_0 with the camera's optical axis pointing to the floor perpendicularly.

Step. 2 Place a calibration pattern O right under the camera, take an image of it, extract feature points from the image, and find the coordinates of them.

Step. 3 Measure the real-world coordinates of the points in the calibration pattern, which correspond to the extracted feature points in the image.

Step. 4 (*Quadrilateral mapping*) Use a *quadrilateral mapping* technique to construct a *basic space-mapping tabl*e T, which maps each image coordinate pair (u_i, v_j) to a real-world coordinate pair (x_{ij}, y_{ij}), that is, T: (u_i, v_j) → (x_{ij}, y_{ij}).

Stage 2. Modifying basic table for a new environment (see Fig 1(b)).

Step. 5 (*Ceiling-height adaptation*) If the in-field camera setup to be carried out includes just a change of the ceiling height H_0, perform the following operations to modify the basic table T; else, go to the next step.

5.1 Affix the camera to the ceiling, measure the ceiling height with respective to the floor, and denote it as H_1.

5.2 Modify table T to construct a new one with H_1 as input by a technique of *ceiling height adaptation* described later, and go to Step 7.

Step. 6 (*Camera-orientation adaptation*) Perform the following operations to modify the basic table T.

6.1 Affix the camera to the ceiling, measure the ceiling height and the camera's orientation, and denote them as L and θ.

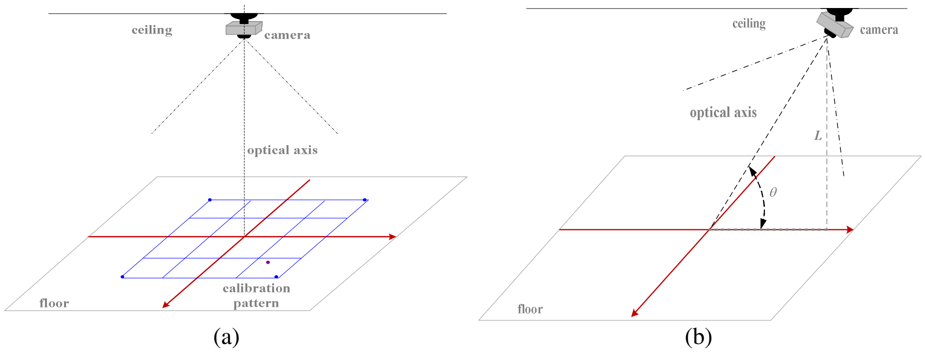
6.2 Modify table T with L and θ as input to be T_1 by a technique of *camera orientation adaptation* described later.

Step. 7 (*Location estimation*) Locate an object B in the real-world space using T_1 in the following way.

7.1 Acquired an image I of B with the camera.

7.2 Detect B in I and find a feature point p on it with coordinates (u, v).

7.3 Use (u, v) to look up T to get the real-world coordinates (x, y) of the real-world point P corresponding to p as the desired object location.



(a) (b)

Fig. 1. Illustration of camera setup. (a) Construction of space-mapping table in Stage 1 of proposed method. (b) Camera orientation change with a tilt angle of θ.

3 Basic Space-Mapping Table Construction and Modifications

The quadrilateral mapping technique mentioned in Step 4 of Algorithm 1 as proposed in this study constructs a space-mapping table T by two steps: finding pairs of corresponding quadrilaterals in the calibration pattern in the image and in the real world, followed by transformations of the image and real-world coordinates of corresponding points within the quadrilaterals based on bilinear interpolation, as illustrated by Figs. 2 and 3. The details are omitted due to the page limit.

After the basic table is obtained with the camera affixed to a ceiling at a certain height H_0 with respect to a floor F_0, if the camera is used later at a different ceiling height H_1 with respect to a second floor F_1, then the table is no more applicable and table content modification is necessary, which we call *ceiling-height adaptation* in Step 5 in Algorithm 1. To do this, first note that an image point p is formed in principle by *any* of the real-world points which all lie on a light ray R going into the camera's lens and then onto the image plane. As illustrated in Fig. 4(a), suppose that this light ray R intersects both the floor F_0 at P_0 and the floor F_1 at P_1. If the image coordinates of p are (u, v), then the real-world coordinates (x_0, y_0) in the basic table corresponding to (u, v) actually are those of P_0 on F_0 instead of being the desired ones, (x_1, y_1), of P_1 on F_1. To correct this error, we derive first the following equalities according to the concept of side proportionality in a triangle:

$$x_1 = x_0 \frac{H_1}{H_0}; \qquad y_1 = y_0 \frac{H_1}{H_0}. \tag{1}$$

That is, the table lookup result (x_0, y_0) corresponding to the image coordinates (u, v) of a real-world point P_1 on F_1 should be magnified in proportion to H_1/H_0 to be (x_1, y_1) as the desired result. Note that we assume here the real-world coordinate system x-y-z is set up at the camera's lens center and the optical axis as the z-axis, and that the location of object point P described by (x_1, y_1) is measured with respect to this system.

Now, assume that the camera is affixed to the ceiling with a tilt angle of θ and a height of L with respect to floor F_1, as shown in Fig. 4(b). Here, the location of object point P_1 on F_1 to be estimated is specified by the real-world coordinates (x_1, y_1) with

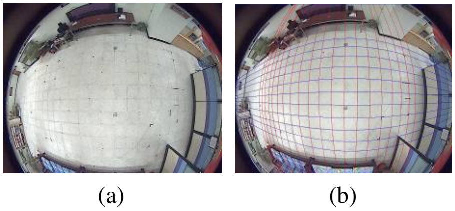


Fig. 2. Illustration of quadrilateral extraction using a grid pattern on floor. (a) An image of the grid pattern. (b) The lines approximating the grid lines.

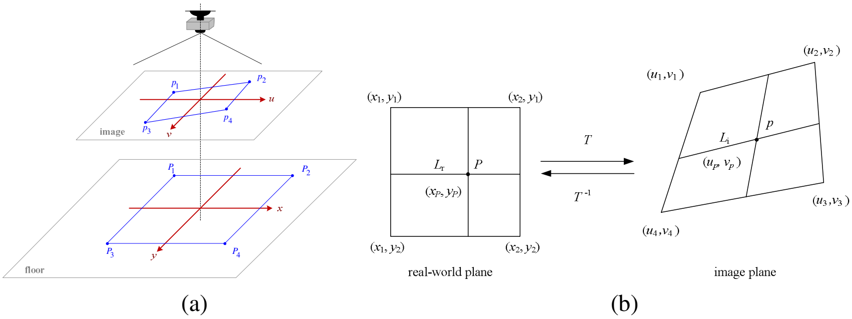


Fig. 3. Quadrilateral mapping. (a) Mapping of corresponding quadrilaterals in image and in calibration pattern. (b) Location estimation of a space point by inverse bilinear interpolation.

respect to the downward projection point O of the camera's lens center onto F_1, where the x-axis is assumed to be coincident with the projection of the camera's optical axis on F_1. Let the coordinates of P_1 in the acquired image be (u, v). Again the basic table is inapplicable here; the table lookup result, the real-world coordinates (x_0, y_0), are actually those of a real-world point on a floor F_0 at a distance H_0 to the camera's lens center, instead of being the desired real-world coordinates (x_1, y_1) of P_1 on F_1. Again, table modification is necessary here, which is called *camera orientation adaptation* in Step 6 of Algorithm 1.

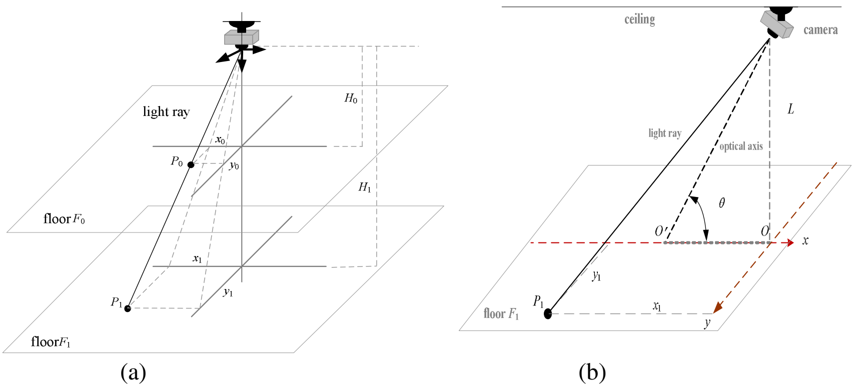
To correct the values (x_0, y_0) into (x_1, y_1), we rotate F_1 through an angle of $90^\circ - \theta$ with P_1 as the rotation pivot point, such that the resulting plane F_1' becomes perpendicular to the camera's optical axis and the lateral view of the rotation result seen from the positive y-axis direction becomes the one shown in Fig. 5. The original floor F_0 is also shown in the figure. Assume that the distance of P_1 on F_1' to the camera's optical axis is x'. Then, according to the concept of side proportionality again, we have

$$\frac{x_0}{x'} = \frac{H_0}{H_1}. \tag{2}$$

Also, by geometry and trigonometry we have

$$\sin\theta = \frac{x'}{M}; \tag{3}$$

$$\sin\theta = \frac{L}{N + H_0}; \tag{4}$$



(a) (b)

Fig. 4. (a) Use of side proportionality to compute coordinates of point P_1 on a floor F_1 with a ceiling height H_1. (b) A tilted camera with angle θ with respect to the x-axis of the real-world coordinate system.

$$\cos\theta = \frac{x_1 - M}{N + H_0}\,; \tag{5}$$

$$\cos\theta = \frac{H_1 - (N + H_0)}{M}\,. \tag{6}$$

From (4) and (5), we get $N + H_0 = L/\sin\theta = (x_1 - M)/\cos\theta$, or equivalently,

$$(x_1 - M)\sin\theta = L\cos\theta. \tag{7}$$

Also, from (2) and (3), we get $x_0M/\sin\theta = H_0/H_1$, or equivalently,

$$H_1 = \frac{H_0 \sin\theta}{x_0 M}\,. \tag{8}$$

From (4), (6) and (8), we get

$$M = \frac{L}{\sin\theta} \times \frac{x_0}{H_0 \sin\theta - x_0 \cos\theta}\,. \tag{9}$$

And from (7) and (9), we get

$$x_1 = L \times \frac{H_0 \cos\theta + x_0 \sin\theta}{H_0 \sin\theta - x_0 \cos\theta}\,. \tag{10}$$

On the other hand, because the x-axis on F_1 is assumed to be coincident with the projection of the camera's optical axis on F_1 and because the rotation of F_1 into F_1' is pivoted in the y-direction, we have $y' = y_1$. Also, according to Eqs. (1) we have $y'/y_0 = H_1/H_0 = x'/x_0$. Therefore, $y_1 = y' = y_0(x'/x_0)$, from which and (3) and (9), we get

$$y_1 = L \times \frac{y_0}{H_0 \sin\theta - x_0 \cos\theta}\,. \tag{11}$$

4 Experimental Results

A series of experiments have been conducted to test the precision of the proposed method for object location estimation. The fish-eye camera used in the experiments is

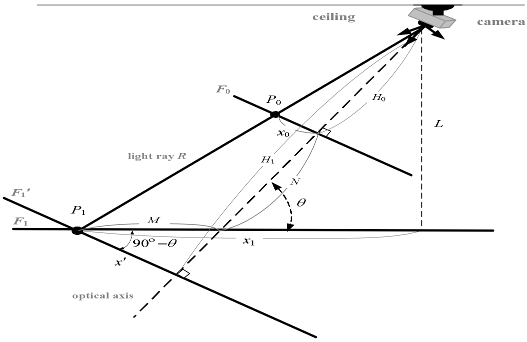


Fig. 5. Lateral view (from the positive y-axis direction) of rotation result of floor F_1 in Fig. 4(b) through an angle of $90^\text{o} - \theta$ with P_1 as the rotation pivot point

shown in Fig. 6(a), which was attached to a rotator connected to a rod with an adjustable length. The camera can so be tilted arbitrarily and raised to any height. An image taken with the camera looking downward is shown in Fig. 2. We show additionally here three images (Fig. 6(b) through 6(d)) taken with the camera in three distinct setups, which are used in our experiments: (1) looking downward at the height of 200cm; (2) looking downward at the height of 250cm; (3) tilted for the angle of 50^o at the height of 200cm. The images are all of the resolution of 1280×1024.

Case (1) is regarded as the *original* camera setup configuration used for building a basic space-mapping table. After the image of Fig. 6(b) was taken with the downward-looking camera at the height 200cm, all the grid points in the image are extracted to get their image coordinates, forming a set denoted by I_c. Also, the real-world coordinates of each grid point are measured manually to form a set denoted by W_c. The two sets I_c and W_c of coordinate data are then used to construct a basic space-mapping table T by the process described in Section 3. To test the precision of the constructed table T, nine non-grid points among the grid ones, which also appear in Fig. 6(b), were selected and their image coordinates collected to form a set I_c'. Also, the real-world coordinates of these non-grid points are measured manually to form another set W_c'. The set I_c' then is used to obtain their corresponding real-world coordinates by table lookup using T, forming a set denoted by W_c''. Finally, the two sets W_c' and W_c'' are compared and two types of error ratio measures are defined to compute the similarity between them: (1) type 1 --- location error ratio with respect to the *distance from the real-world point to the camera's lens center*:

$$\textit{location error ratio} = \frac{\sqrt{(real\ x_i - estimated\ x_i)^2 + (real\ y_i - estimated\ y_i)^2}}{\sqrt{real\ x_i^{\ 2} + real\ y_i^{\ 2} + L^2}}$$

where *real* x_i and *real* y_i are data in W_c' and *estimated* x_i and *estimated* y_i are data in W_c''; (2) type 2 --- location error *ratio* with respect to the *effective field of view of the camera*:

$$\textit{location error ratio} = \frac{\sqrt{(real\ x_i - estimated\ x_i)^2 + (real\ y_i - estimated\ y_i)^2}}{\text{radius of effective camera's field of view}}.$$

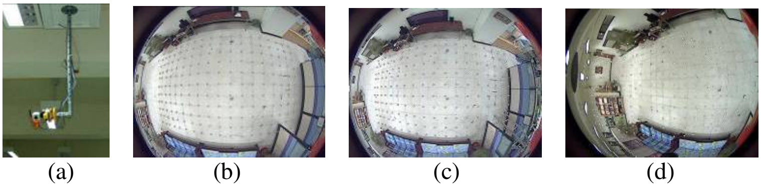


Fig. 6. Fish-eye camera and images used for experiments. (a) The camera can be tilted and lifted arbitrarily. (b)-(d) Images taken with the camera looking downward at heights 200cm and 250cm, and tilted for 50^o at height 200cm.

Table 1. Error ratios with camera looking downward at ceiling height 200cm

real x (cm)	*estimated x (cm)*	*real y (cm)*	*estimated y (cm)*	*distance to origin (cm)*	*type-1 error ratio*	*type-2 error ratio*
-7	-8	-24	-23	25	0.7%	0.4%
-37	-36	36	36	52	0.5%	0.3%
-20	-20	96	94	98	0.9%	0.6%
-45	-44	-107	-106	116	0.6%	0.4%
-111	-112	-55	-56	124	0.6%	0.4%
-140	-140	62	60	153	0.8%	0.6%
-229	-228	-101	-104	250	1.0%	1.0%
-253	-257	76	82	264	2.2%	2.3%
-320	-317	-15	-15	320	0.8%	0.9%

Table 2. Error ratios with camera looking downward at ceiling height 250cm

-7	-9	-24	-23	25	0.9%	0.7%
-37	-38	36	37	52	0.6%	0.4%
-20	-21	96	94	98	0.8%	0.7%
-45	-48	-107	-109	116	1.3%	1.1%
-111	-117	-55	-57	124	2.3%	2.0%
-140	-145	62	62	153	1.7%	1.6%
-229	-238	-101	-110	250	3.6%	4.0%
-253	-264	76	80	264	3.2%	3.7%
-320	-335	-15	-14	320	3.9%	4.7%

Table 3. Error ratios with camera tilted for 50^o at ceiling height 200cm

real x (cm)	*estimated x (cm)*	*real y (cm)*	*estimated y (cm)*	*distance to origin (cm)*	*type-1 error ratio*	*type-2 error ratio*
-7	-4	-24	-20	25	2.5%	1.6%
-37	-34	36	36	52	1.5%	0.9%
-20	-19	96	93	98	1.4%	1.0%
-45	-39	-107	-103	116	3.1%	2.3%
-111	-106	-55	-59	124	2.7%	2.0%
-140	-138	62	57	153	2.1%	1.7%
-229	-234	-101	-116	250	4.9%	4.9%
-253	-265	76	75	264	3.6%	3.8%
-320	-342	-15	-25	320	6.4%	7.5%

The computed results for the two types of error ratios are summarized as a table shown in Table 1, from which we can see the ratios are all smaller than 5% which is practical for object location estimation applications like indoor vehicle guidance.

For Case (2), the camera, still looking downward, was affixed at a different height 250cm and the previously-mentioned process of error ratio computation was repeated after the proposed method was applied to the image of Fig. 6(c). The results were again summarized as a table shown in Table 2, from which we can see the ratios are all smaller than 5% as well. Similarly, for Case (3) where the camera was affixed at the height 200cm and tilted for 50^o, the error ratio table constructed for the image of Fig. 6(d) is shown in Table 3, from which we see the ratios are *not* all smaller than 5% this time; some are larger (6.4% and 7.5% for the last row in the table). The reason for this phenomenon is that the object point dealt with is located at (−320, −15) which is quite far away from the center of the image, falls within a distorted-shaped quadrilateral, and so incurs a larger error in the quadrilateral mapping process.

5 Conclusions

A general space-mapping method for object location estimation by modifying the basic space-mapping table for camera setup change adaptation has been proposed. The method does not require conventional camera calibration, and is general for any camera type. The method estimates the location of an object by mapping the image coordinates of object points to the real-world coordinates of the points using a space-mapping table. An algorithm is designed to construct the table, which consists of two stages, with the first for constructing a basic table using bilinear interpolation and the second for modifying the table to adapt it to changes of camera heights and orientations, which often occur in different application environments. Experimental results show that the method yields results with error ratios smaller than 5% in most cases, meaning the practicality of the method for various applications.

References

1. Chou, H.L., Tsai, W.H.: A new approach to robot location by house corners. Pattern Recognition 19(6), 439–451 (1986)
2. Betke, M., Gurvits, L.: Mobile robot localization using landmarks. IEEE Transactions on Robotics and Automation 13(2), 251–263 (1997)
3. Aider, O.A., Hoppenot, P., Colle, E.: A model-based method for indoor mobile robot localization using monocular vision and straight-line correspondences. Robotics and Autonomous Systems 52(2-3), 229–246 (2005)
4. Hemayed, E.E.: A survey of camera self-calibration. In: Proceedings of IEEE Conference on Advanced Video and Signal Based Surveillance, Miami, FL, USA, pp. 351–357 (2003)
5. Yang, Z.F., Tsai, W.H.: Viewing corridors as right parallelepipeds for vision-based vehicle localization. IEEE Trans. Industrial Electronics 46(3), 653–661 (1999)
6. Takeshita, T., Tomizawa, T., Ohya, A.: A house cleaning robot system – path indication and position estimation using ceiling camera. In: Proceedings of International Joint Conference on SICE-ICASE, Busan, Korea, pp. 2653–2656 (2006)
7. Wang, Y.T., Tsai, W.H.: Indoor security patrolling with intruding person detection and following capabilities by vision-based autonomous vehicle navigation. In: Proceedings of 2006 International Computer Symposium (ICS 2006), Taipei, Taiwan (2006)
8. Jeng, S.W., Tsai, W.H.: Using pano-mapping tables to unwarping of omni-images into panoramic and perspective-view Images. IET Image Pro. 1(2), 149–155 (2007)
9. Chen, H.C., Tsai, W.H.: Optimal security patrolling by multiple vision-based autonomous vehicles with omni-monitoring from the ceiling. In: Proceedings of 2008 International Computer Symposium, Taipei, Taiwan (2008)

A Novel Method for Lateral Vehicle Localization by Omni-Cameras for Car Driving Assistance*

Chih-Jen Wu[1] and Wen-Hsiang Tsai[1,2]

[1] Institute of Computer Science and Engineering, National Chiao Tung University, Taiwan
gis91813@cis.nctu.edu.tw
[2] Department of Information Communication, Asia University, Taiwan
whtsai@cis.nctu.edu.tw

Abstract. A lateral vehicle localization method by omni-image analysis is proposed for car driving assistance. The method estimates analytically the position and orientation of a lateral vehicle by utilizing the geometric properties of a circular-shaped wheel image of the lateral car taken by a single omni-camera with a hyperboloidal-shaped mirror. Analytical solutions are made possible for fast computation by a special arrangement of affixing the omni-camera on the frontal car bumper at the height of the wheel. Experimental results showing good data estimation precision are included to prove the feasibility of the proposed method.

Keywords: vehicle localization, hyperboloidal-shaped mirror, omni-camera, circle, car wheel.

1 Introduction

Car driving assistance using traditional cameras has been studied intensively [1-4]. Recently, omni-cameras with wider views become popular. They are more suitable for car driving assistance because fewer cameras need be equipped. For example, Lai and Tsai [3] affixed a traditional camera on the right-frontal side of a *host car* to take the image of a *lateral car*. To acquire a full frontal view, two more traditional cameras should be used. Instead, one frontal omni-camera is sufficient. Additionally, car wheels are circular-shaped, providing geometric hints for lateral car localization [3]. However, when a circle appears in an omni-image, it becomes irregular in shape and cannot be described mathematically [5], leading to difficulty of extending the existing vehicle localization methods for omni-images.

In this study, we try to solve this problem. The omni-camera is equipped on the frontal bumper *at the height of the wheel* so that the mathematics involved in circular-shape image analysis becomes maneuverable to get analytic solutions for fast computation. In the following, the proposed method is described in Section 2, followed by some experimental results in Section 3 and conclusions in Section 4.

* This work was supported financially by the Ministry of Economic Affairs under Project No. MOEA 97-EC-17-A-02-S1-032 in Technology Development Program for Academia.

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 403–410, 2009.

2 Lateral Car Localization by Frontal Omni-Camera

The basic idea of the proposed method is to utilize the geometric properties of a circular-shaped wheel image of the lateral car taken by a single omni-camera to estimate the position and orientation of the lateral car with respect to the host car. The omni-camera, affixed to the frontal bumper of the host car at the height of the wheel, includes a hyperboloidal-shaped mirror. Also, the optical axis of the camera is set to be horizontal to the ground plane. Such an arrangement of the camera makes the resulting irregular shape of the wheel in the omni-image to be extractable as an ellipse using the Hough transform, as proved in [5]. More details are described by the following algorithm.

Algorithm 1. ***Lateral car localization by a frontal omni-camera on a host car.***

Step 1. Affix an omni-camera to the frontal bumper of the host car at the height of the wheel center with the camera's optical axis adjusted to be horizontal to the ground plane and pointing to the frontal direction of the host car.

Step 2. Take an image of a wheel of the lateral car and find out the vertical height h of the wheel in the image.

Step 3. With the radius of the wheel and the value of h as input, estimate the position of the lateral car (details described in Section 2.1).

Step 4. Find out the farthest and the closest points, I_f and I_c, of the wheel in the image with respect to the image center.

Step 5. With I_f and I_c as input, derive the direction of the lateral car with respect to the host car (details described in Section 2.2).

2.1 Estimation of Lateral Car Position Using Rotational Invariance Property

The coordinate systems involved in an omni-camera system, including a traditional perspective camera and a hyperboloidal-shaped mirror, are depicted in Fig. 1(a), where the omni-camera and the image coordinates are specified by (X, Y, Z), and (u, v), respectively. The perspective camera and the mirror are properly aligned, as assumed, so that the omni-camera becomes a *single-viewpoint* system, and that the optical axis of the perspective camera coincides with the mirror axis which is the line going through the mirror center and perpendicular to the mirror base.

The middle point between the camera lens center O$_l$ and the mirror focus point O$_m$ is taken to be the origin O$_a$ of the omni-camera coordinate system. The hyperboloidal mirror shape may so be described by

$$\frac{r^2}{a^2}-\frac{Z^2}{b^2}=-1, \qquad r=\sqrt{X^2+Y^2} \tag{1}$$

and O$_m$ is located at (0, 0, $-c$) and O$_l$ at (0, 0, $+c$) in the camera coordinate system where $c=\sqrt{a^2+b^2}$. The relationship between (u, v) and (X, Y, Z) may be described [5] by

$$u=\frac{Xf(b^2-c^2)}{(b^2+c^2)(Z-c)-2bc\sqrt{(Z-c)^2+X^2+Y^2}};$$

$$v = \frac{Yf(b^2 - c^2)}{(b^2 + c^2)(Z - c) - 2bc\sqrt{(Z - c)^2 + X^2 + Y^2}} \tag{2}$$

where f is the camera's focal length, and b, c, and f are parameters assumed to be known in advance.

Also, as illustrated in Fig. 1(b), the omni-camera is affixed to the car bumper with the Z-axis of the omni-camera adjusted to be at the height of the wheel center, the negative Z-axis directed to the car driving direction, and the Y-axis set perpendicular to the ground surface. A wheel coordinate system x-y-z is defined on the left-frontal wheel of the lateral car with its origin O_w being the wheel center and its x-y plane being the wheel plane. The orientation of the wheel plane is denoted by θ with $\theta = 0^{\circ}$ meaning that the lateral car moves in parallel. The wheel's radius is assumed to be R. Then, defining (X_c, Y_c, Z_c) as the wheel center's coordinates in the omni-camera coordinate system, we get

$$Y_c = 0; \tag{3}$$

$$Y = y. \tag{4}$$

From the above omni-camera geometry, it is not to difficult to figure out the validity of the so-called *rotational invariance* property, which means that the angle of an incoming light ray formed by a space point at coordinates (X, Y, Z) onto the mirror surface in the omni-camera coordinate system is identical to the angle of the corresponding image point at coordinates (u, v) in the image coordinate system, leading to the following equality:

$$\frac{v}{u} = \frac{Y}{X}. \tag{5}$$

The lateral car localization problem now is to derive the wheel position and orientation parameters X_c, Z_c, and θ of the lateral car in the omni-camera coordinate system. First, define P_1 through P_4 as the four extreme points on the wheel circle so that the segments $\overline{P_1P_2}$ and $\overline{P_3P_4}$ are perpendicular and parallel to the ground surface, respectively, as illustrated by Fig. 2. Obviously, P_1 and P_2 are at (X_c, $+R$, Z_c) and (X_c, $-R$, Z_c), respectively. Next, it can be figured out that the Y-coordinates of P_3 and P_4 are equal to Y_c, which is zero, because the wheel center O_w at (X_c, Y_c, Z_c) is at the height of the omni-camera coordinate system origin O_a at coordinates (0, 0, 0), as assumed. Denote the image point corresponding to P_i as I_i and its coordinates as (u_i, v_i), i = 1, 2, 3, 4. Applying (5) to I_1 and I_2, we get

$$\frac{v_1}{u_1} = \frac{R}{X_c}; \quad \frac{v_2}{u_2} = \frac{-R}{X_c}. \tag{6}$$

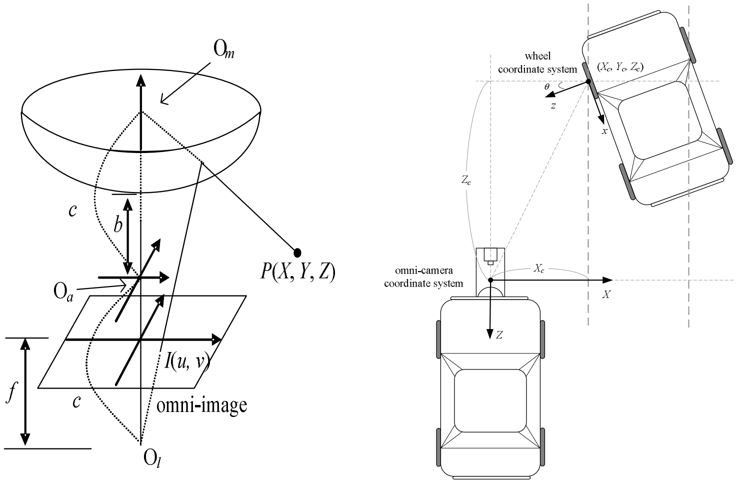
Also, from (2) and the omni-camera coordinates of P_1 and P_2, we get

$$u_1 = u_2. \tag{7}$$

Combining (6) and (7), we get the solution for X_c as

$$X_c = \frac{2R}{v_1 - v_2} u_1. \tag{8}$$

Note that $v_1 - v_2$ is just the value h mentioned in Step 3 of Algorithm 1.



(a) (b)

Fig. 1. Relative coordinate systems. (a) Omni-camera and image coordinate systems. (b) Omni-camera and wheel coordinate systems.

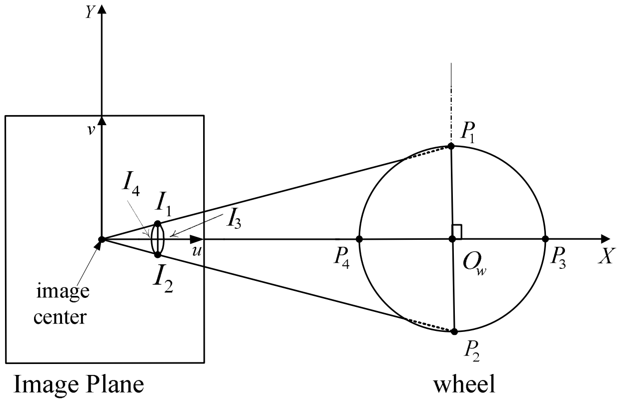


Fig. 2. Definition of corresponding image and space points

To derive Z_c, let

$$Z_c' = Z_c - c. \tag{9}$$

Then, (2) for I_1 may be transformed into

$$[u_1^2(b^2 + c^2)^2 - u_1^2(-2bc)^2](Z_c')^2 - [2u_1X_cf(b^2 - c^2)(b^2 + c^2)](Z_c') \\ + [f^2(b^2 - c^2)^2X_c^2 - u_1^2(-2bc)^2X_c^2] = 0$$

which leads to

$$A(Z_c')^2 + B(Z_c') + C = 0 \tag{10}$$

where

$$\begin{aligned} A &= u_1^{\ 2}(b^2 - c^2)^2; \\ B &= -2u_1 X_c f(b^2 - c^2)(b^2 + c^2); \\ C &= X_c^{\ 2}[f^2(b^2 - c^2)^2 - 4b^2c^2u_1^{\ 2}]. \end{aligned} \tag{11}$$

So, we get

$$Z_c' = \frac{-B \pm \sqrt{B^2 - 4AC}}{2A} \tag{12}$$

which may be simplified to be

$$Z_c' = \frac{X_c[f(b^2 + c^2) \pm 2bc\sqrt{f^2 + u_1^{\ 2}}]}{u_1(b^2 - c^2)}. \tag{13}$$

With the solution for X_c in (8), $Z_c' = Z_c - c$ in (9), and the equation of (13) above, we get finally the solution for Z_c as

$$\begin{aligned} Z_c &= Z_c' + c \\ &= \frac{X_c[f(b^2 + c^2) \pm 2bc\sqrt{f^2 + u_1^{\ 2}}]}{u_1(b^2 - c^2)} + c \\ &= \frac{2R[f(b^2 + c^2) \pm 2bc\sqrt{f^2 + u_1^{\ 2}}]}{(v_1 - v_2)(b^2 - c^2)} + c. \end{aligned} \tag{14}$$

The sign (+ or –) in (14) may be decided experimentally.

2.2 Estimation of Lateral Car Orientation Using Wheel Shape Information

Now, we want to use the image coordinates (u_3, v_3) and (u_4, v_4) of I_3 and I_4 (denoted as I_f and I_c respectively in Step 4 of Algorithm 1) to derive the wheel orientation θ based on the values of X_c and Z_c obtained previously. First, according to (2) and because $Y_3 = Y_4 = Y_\mathrm{c} = 0$, we get $v_3 = v_4 = 0$ and

$$u_3 = \frac{X_3 f(b^2 - c^2)}{(b^2 + c^2)(Z_3 - c) - 2bc\sqrt{(Z_3 - c)^2 + X_3^{\ 2} + Y_3^{\ 2}}}; \tag{15}$$

$$u_4 = \frac{X_4 f(b^2 - c^2)}{(b^2 + c^2)(Z_4 - c) - 2bc\sqrt{(Z_4 - c)^2 + X_4^{\ 2} + Y_4^{\ 2}}}. \tag{16}$$

Also, define

$$Z_3' = \mathrm{Z}_3 - c; \tag{17}$$

$$Z_4' = \mathrm{Z}_4 - c. \tag{18}$$

Using (15) and (16) and through a similar process to that for deriving (13), we get

$$Z_3' = \frac{X_3[f(b^2+c^2) \pm 2bc\sqrt{f^2+u_3^{\,2}}]}{u_3(b^2-c^2)}; \tag{19}$$

$$Z_4' = \frac{X_4[f(b^2+c^2) \pm 2bc\sqrt{f^2+u_4^{\,2}}]}{u_4(b^2-c^2)}. \tag{20}$$

Furthermore, with the middle point of $\overline{P_3P_4}$ as the wheel center O$_w$, we get

$$X_4 = 2X_c - X_3; \tag{21}$$

$$Z_4 = 2Z_c - Z_3. \tag{22}$$

Combining (17), (18) and (22), we have

$$Z_4' = 2Z_c - Z_3' - 2c. \tag{23}$$

Also, (19) and (20) may be transformed into

$$Z_3' = X_3A_3; \tag{24}$$

$$Z_4' = X_4A_4 \tag{25}$$

where

$$A_3 = \frac{[f(b^2+c^2) \pm 2bc\sqrt{f^2+u_3^{\,2}}]}{u_3(b^2-c^2)}; \tag{26}$$

$$A_4 = \frac{[f(b^2+c^2) \pm 2bc\sqrt{f^2+u_4^{\,2}}]}{u_4(b^2-c^2)}. \tag{27}$$

Combining (21) and (23) through (25), we get

$$X_3 = 2(Z_c - X_cA_4 - c)/(A_3 - A_4). \tag{28}$$

And from (17), (24), and (28), we get

$$Z_3 = (2Z_cA_3 - 2X_cA_3A_4 - cA_3 - cA_4)/(A_3 - A_4). \tag{29}$$

And from (22) and (29), we get

$$Z_4 = (2X_cA_3A_4 + cA_3 + cA_4 - 2Z_cA_4)/(A_3 - A_4). \tag{30}$$

Accordingly, from (25) we get

$$X_4 = Z_4'/A_4 = (Z_4 - c)/A_4$$

$$= (2X_cA_3A_4 + cA_3 + cA_4 - 2Z_cA_4 - c)/[(A_3 - A_4)A_4]. \tag{31}$$

With (28) through (31), we finally get the desired result

$$\theta = \tan^{-1}(\frac{X_3 - X_4}{Z_3 - Z_4}).$$

3 Experimental Results

In our experiments a set of real location data of a lateral car in different postures were measured before corresponding images were taken to estimate the posture parameters X_c, Z_c, and θ using the previously-derived equations. An example of the acquired images is shown in Fig. 3, in which detection of a wheel shape as an ellipse is also shown. Some estimation results are shown in Table 1. The error for X_c or Z_c is computed as the ratio of the difference between the estimated value and the real one with respect to the real value. And the angle error for θ is computed similarly but with respect to 180^{o} which is the angle range of the lateral car. The table shows that the estimated values of X_c and Y_c are within 5% errors which are good enough for practical applications. But some angle errors are larger. The reason is that the wheel size in the images of these cases appeared to be small, so that the estimated angle θ is sensitive to the width of the horizontal wheel diameter $\overline{P_3P_4}$.

Moreover, in Section 2.1 the radius of the wheel should be known in advance for estimating the vehicle location. However, in real cases of applying the proposed vehicle localization method, the type of the wheel on the vehicle is unknown in advance. A solution is to assume an average wheel radius, which is 20.75cm according to our measurement of a lot of car wheel radiuses. But this way will introduce errors in the estimated position values. Therefore, a simulation experiment was conducted to test

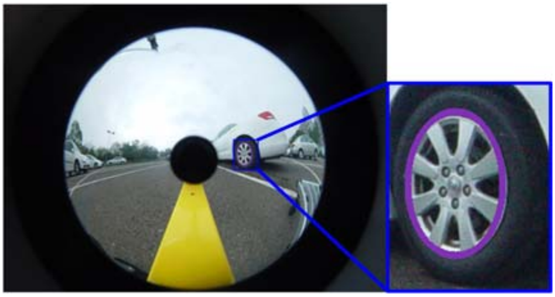


Fig. 3. A lateral car image with wheel shape detected as an elliptical shape

Table 1. Lateral car location estimation results

real X_c (cm)	*estimated X_c (cm)*	*error ratio of X_c*	*real Z_c (cm)*	*estimated Z_c (cm)*	*error ratio of Z_c*	*real θ (degree)*	*estimated θ (degree)*	*angle error (degree)*
-104.3	-99.7	4.4%	-40.9	-39.3	4.0%	-5	-6	0.6%
-390.4	-382.4	2.2%	-87.2	-87.7	0.6%	15	13	1.1%
-464.9	-478.0	4.8%	-494.9	-502.7	1.6%	22	38	8.9%
-137.7	-138.4	0.7%	-159.9	-166.6	4.2%	9	21	6.7%
-407.0	-419.0	4.3%	-115.2	-112.5	2.3%	33	22	6.1%
608.0	606.8	0.6%	-235.5	-240.3	2.0%	-82	-73	5.0%
average error		2.8%			2.5%			4.7%

Table 2. Simulation results of estimating lateral car position using a fixed wheel radius value 20.75 cm

input radius of wheel (cm)	19.75		20.41		21.08		21.75	
position coordinates	*X*	*Z*	*X*	*Z*	*X*	Z	*X*	*Z*
input real values (cm)	400	100	400	100	400	100	400	100
estimated values (cm)	419.24	104.81	405.55	101.39	392.73	98.18	380.69	95.17
error ratio	4.81%	4.81%	1.39%	1.39%	−1.82%	−1.82%	−4.82%	−4.82%

whether the errors are tolerable or not. For this, first we project the wheel shape of a lateral car onto the image plane using a set of different real wheel radius parameters (ranging from 19.75cm to 21.75cm as measured by us). Then the proposed method was applied to estimate the position of the lateral car, under the assumption that the radius of the wheel of the car is of the above-mentioned average value 20.75cm. The results are shown in Table 2. The average error rates of the position parameters are all smaller than 5%, so the use of a fixed wheel radius in deriving the lateral car location is considered feasible in practice.

4 Conclusions

A lateral vehicle localization method by the use of a single frontal omni-camera has been proposed. The basic concept is to affix a single omni-camera on the bumper at the height of the wheel so that the mathematics involved in the analysis of the irregular shape of the circular wheel in the image becomes maneuverable, leading to the possibility of deriving analytic solutions. Experimental results show that most location estimation results are with error ratios smaller than 6%, which means that the proposed method is feasible for practical applications.

References

1. Lin, H.H., Lin, J.H.: A Visual Positioning System for Vehicle or Mobile Robot Navigation. IEICE Transactions on Information and Systems E89-D(7), 2109–2116 (2006)
2. Sakurai, K., Kyo, S., Okazaki, S.: Overtaking Vehicle Detection Method and Its Implementation Using IMAPCAR Highly Parallel Image Processor. IEICE Transactions on Information and Systems E91-D(7), 1899–1905 (2008)
3. Lai, C.C., Tsai, W.H.: Estimation of Moving Vehicle Locations Using Wheel Shape Information in Single 2-D Lateral Vehicle Images by 3-D Computer Vision Techniques. Robotics and Computer Integrated Manufacturing 15, 111–120 (1999)
4. Cao, Y., Renfrew, A., Cook, P.: Vehicle Motion Analysis Based on a Monocular Vision System. In: Proceedings of 2008 Road Transport Information and Control and ITS Conference, Manchester, UK, pp. 1-6 (2008)
5. Wu, C.J., Tsai, W.H.: Location Estimation for Indoor Autonomous Vehicle Navigation by Omni-Directional Vision Using Circular Landmarks on Ceilings. Robotics and Autonomous Systems 57(5), 546–555 (2009)

Abnormal Event Analysis Using Patching Matching and Concentric Features

Jun-Wei Hsieh[1,2], Sin-Yu Chen[1], and Chao-Hong Chiang[1]

[1] Department of Electrical Engineering
Yuan Ze University,
135 Yuan-Tung Road, Chung-Li 320, Taiwan, R.O.C.
[2] Department of Computer Science and Engineering
National Taiwan Ocean University
2 Pei-Ning Road, Keelung 202-24, Taiwan, R.O.C.

Abstract. This paper proposes a novel patch-based approach for abnormal event detection from a mobile camera using concentric features. It is very different from traditional methods which require the cameras being static for well foreground object detection. Two stages are included in this system i.e., training and detection, for scene representation and exceptional change detection of important objects like paintings or antiques. Firstly, at the training stage, a novel scene representation scheme is proposed for large-scale surveillance using a set of corners and key frames. Then, at the detection stage, a novel patch matching scheme is proposed for efficient scene searching and comparison. The scheme reduces the time complexity of matching not only from search space but also feature dimension in similarity matching. Thus, desired scenes can be obtained extremely fast. After that, a spider-web structure is proposed for missing object detection even though there are large camera movements between any two adjacent frames. Experimental results prove that our proposed system is efficient, robust, and superior in missing object detection and abnormal event analysis.

Keywords: Video surveillance, patch clustering, concentric features, and scene clustering.

1 Introduction

Video surveillance [1]-[2] is a system which analyzes different security events directly from videos. For example, the task of missing object detection can be used for security monitoring, crime detection, and anti-terrorist surveillance. In most surveillance systems, the camera should be fixed to the background so that foreground objects can be extracted through a subtraction technqiue. For example, in [3], Kim et al. built a codebook model to extract foreground objects directly from videos. Stringa [4] used a key frame extraction technique to locate moving objects and then recognized suspicious objects according to their positions, moments, and shape features. In [5], Foresti et al. used a long-term change detection algorithm to detect abandoned objects

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 411–420, 2009.

and then classified video sequences into four dangerous events. However, a fixed camera is not proper for building a large scale surveillance system due to its limited field-of-views. To develop this wide-scale surveillance system, a mobile camera will be better adopted for monitoring different abnormal event and activities. Thus, in [6], Castelnovi et al. presented a surveillance robot for scene analysis by comparing color similarities or differences between two images. However, this approach requires the two images being well registered.

This paper designs a novel surveillance system which detects exceptional changes of scenes (caused by missing objects or abandoned objects like paintings, antiques, or packages) as abnormal events directly from videos. When a moving camera is used, most of background subtraction techniques [2]-[3] will fail to detect foreground objects due to large changes of scene contents. To tackle this problem, this paper presents a novel patching clustering scheme for scene representation and comparison so that desired missing objects or abnormal events can be well detected in real time even though a mobile camera is used. To represent each observed scene, at the training stage, a clustering scheme is first used for extracting a set of key frames from the surveillance environments. To reduce the effects of lighting changes, each key frame is further represented by a set of corner features. Then, at the detection stage, the task of missing object detection (or abnormal event analysis) will become a scene searching problem. Since each scene includes lots of key frames and each frame is represented by a set of corners, the searching space is very huge. To speed up the searching efficiency, this paper presents a novel patch matching method for searching desired scenes from this huge space in real time. It reduces not only the search space but also the feature space for similarity matching. For the feature space, the proposed method can reduce more than one orders of time complexity in similarity calculation. For pruning the searching space, this paper uses a set of projection kernels (generated from an integral image) to construct a ring structure. It forms a series of weak hypotheses to filter out impossible candidates. Since the kernel function is small and independent to pattern size, the search process can be very efficiently performed for scene searching. After that, a spider-web structure is proposed for missing object detection even though large subtraction errors happen. Experimental results have proved the superiorities of our proposed method in missing object detection and abnormal event analysis.

2 System Overview

Fig. 1 shows the flowchart of our surveillance system in missing object detection and abnormal event analysis. Two stages are included in our proposed approach, i.e., the training and detection stages. At the training phase, we use the scene classification and feature tracking techniques to construct the surveillance scene. At the detection stage, a novel scene matching scheme is then proposed for scene searching and comparison so that different abnormal events caused by missing objects can be well analyzed.

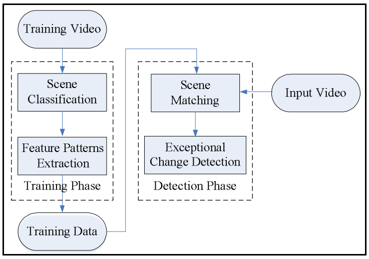


Fig. 1. Flowchart of the system

3 Scene Construction

When a moving camera is used, the task of missing object detection will become very difficult due to large changes of scenes. Thus, a classification scheme is used for classifying the observed scene to different key frames. Then, a set of corners is extracted for representing each key frame so that the observed scene can be well constructed and analyzed.

3.1 Scene Classification

This paper represents a scene with a set of key frames. To select the set of key frames, the similarity between two adjacent frames A and B should be measured. Let $H(A)$ and $H(B)$ denote the color histograms of A and B, respectively. Then, the ordinal distance between A and B is used to measure their dissimilarity, i.e.,

$$D_{ord}\left(H(A),H(B)\right)=\sum_{i=1}^{T-1}\left|\sum_{j=1}^{i}\left(H_j(A)-H_j(B)\right)\right|, \tag{1}$$

where T is the total number of color bins. Suppose K_i is the ith key frame at time t and $I_{t+\tau}$ is the input frame acquired at time $t+\tau$. The (i+1)th key frame can be determined if $D_{ord}\left(H(I_i),H(I_{t+\tau})\right)$ is larger than a threshold and the frame $I_{t+\tau}$ is also the (i+1)th key frame.

3.2 Scene Representation Using Corners

In addition to key frames, a set of corners is also extracted for scene representation. Assume the scene consists of L sub-scenes. For the tth sub-scene, there are N frames between key frames I_t at time t and I_{t+N} at time $t+N$. All the corner features within the N frames are stored for scene representation. However, corners within these N frames are almost overlapped. To more efficiently represent the scene, a correlation technique is used to track each corner across these N frames. Then, the coordinates of all extracted corner are aligned to the coordinates of I_t. After alignment, all the repeated corners are eliminated. Thus, a more efficient representation can be used for scene construction.

4 Detection Phase

The critical stage in missing object detection is scene searching and scene comparison. In what follows, an extremely efficient matching scheme is proposed.

4.1 Pattern Matching

Our proposed pattern matching scheme uses a boosting process and a concentric sampling structure over a pyramid framework to very efficiently match and search image patterns. The time complexity of our scheme is invariant to the dimension of used windows. This method is significantly different from traditional pattern matching schemes using a time-exhausting correlation technique to match patterns. The scheme uses different weak beliefs to filter out impossible candidates through a collection of concentric rotation-invariant sub-samples of the desired pattern. The rotation-invariant sampling is approximated with integral images, and the hierarchy scheme sifts the group of candidates in reduced complexity as a reliable relaxation process. Fig. 2 shows the flowchart for patterns matching.

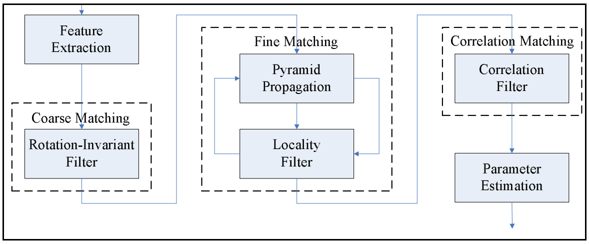


Fig. 2. Flowchart for pattern matching

4.1.1 Feature Selection

For a feature used in our method, we primarily concern its robustness for surviving under different lighting conditions and its low complexity for efficient matching. The pixel intensity can be directly used in our scheme for similarity measurement. However, pixel intensity is not stable when images have illumination changes. Thus, we extract edge feature from image differential spaces for pattern matching since it is suitable for overcoming the problem of lighting changes and verifying the local structures of patterns. Given a grayscale image I, its edge map can be obtained by gradient operations, i. e., $E= \| \nabla \cdot I \|$, where $E(x,y)= | I(x,y)-I(x-1,y) | + | I(x,y)-I(x,y-1) |$.

Assume a two dimension $m \times m$ pattern $p(x,y)$ is to be matched within an input image $I(x,y)$ of size $n \times n$. For one pixel c in the image, we can measure the distance between it and p as follows

$$d^2(c,p) = \sum_{i=-m/2}^{m/2} \sum_{j=-m/2}^{m/2} \left(I(x_c+i, y_c+j) - p(i,j) \right)^2 , \tag{2}$$

where c has the coordinates (x_c,y_c). Clearly, the time complexity of calculating $d^2(c,p)$ is $O(m^2)$. The time complexity can be reduced to $O(m)$ if a kernel function is used based on the concept "integral image" for avoiding lots of redundant calculation.

Given a point *c* in *I*(*x*,*y*), we define its kernel-based sampling function with the radius *ρ* as follows

$$k(c,\rho)=\frac{1}{\|R\|}\int_{p\in R} E(p), \tag{3}$$

where $R=\{p \mid \|p-c\|<\rho\}$ and *E* is the edge map of *I*. The kernel-based sampling function forms the basic statistical measurement in our task for pattern matching. It is noticed that *k*(*c*,*ρ*) is rotation-invariant. When different radii *ρ* are used, a new descriptor can be defined for describing the visual characteristics of *c*, i.e., *k*(*c*,1), *k*(*c*,2),…,*k*(*c*,*m*). Then, a vector *k*(*c*) with *m* elements can be defined for describing *c*, i.e., *k*(*c*)=(*k*(*c*,1), *k*(*c*,2),…,*k*(*c*,*m*)). For the pattern *p*, we also use the similar idea to obtain its descriptor *k*(*p*). Then, the distance between *c* and *p* can be redefined as follows

$$d_k^2(c,p)=\sum_{i=1}^{m} w_i\left(k(c,i)-k(p,i)\right)^2, \tag{4}$$

where w_i is a weight to weight the term *k*(*c*, *i*). Compared with Eq.(2), the time complexity for calculating $d_k^2(c,p)$ is *O*(*m*) which reduces one order of similarity calculation than Eq.(2).

4.1.2 Integral Feature with Rotation Invariance

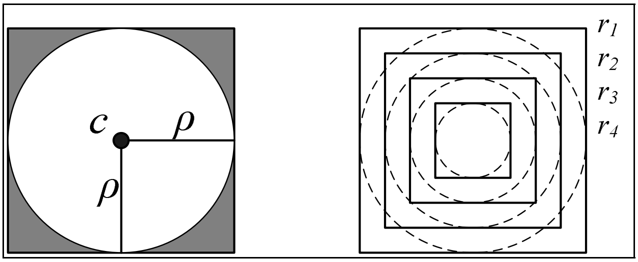
To obtain the descriptor *k*(*c*), we use a circular kernel sampling function to calculate each term *k*(*c*,*ρ*) like Fig. 3(a). Technically, the circular kernel sampling function can be approximated with a square kernel sampling function *π*(*c*,*ρ*) defined as follows

$$\pi(c,\rho)=\frac{1}{\rho^2}\sum_{i=-\rho/2}^{\rho/2}\sum_{j=-\rho/2}^{\rho/2} E(x_c+i,y_c+j). \tag{5}$$

Like Fig. 3 (b), we use *π*(*c*, *ρ*) to approximate *k*(*c*, *ρ*). Since *π*(*c*, *ρ*) can be very efficiently obtained through an integral image structure, we use the descriptor *π*(*c*) to approximate *k*(*c*) for efficient pattern matching.

Given an edge map *E*, its integral image *S*(*x*,*y*) contains the sum of edge points in *E* accumulated from the original (0,0) to the pixel (*x*,*y*), i.e.,

$$S(x,y)=\sum_{i=0}^{x}\sum_{j=0}^{y} E(i,j). \tag{6}$$



(a) (b)

Fig. 3. (a) A circular sampling function. (b) Its squared approximation centered at location *c*.

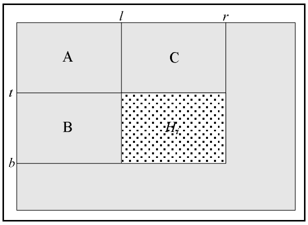


Fig. 4 Calculation of integral image

$S(x, y)$ can be obtained using only one scan over E. Given a rectangle region H_i bounded by (l,t,r,b), its sum of edge magnitudes can be very efficiently achieved by taking advantages of the integral image S. In Fig. 4, $sum(H_i)$ can be easily calculated as follows

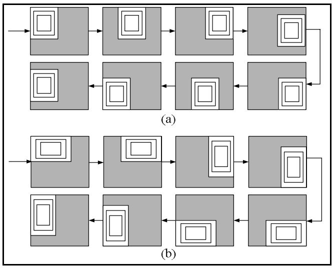
$$\begin{aligned} sum(H_i) &= (A+B+C+H_i)+A-(A+B)-(A+C) \\ &= S(r,b)+S(l,t)-S(l,b)-S(r,t) \end{aligned} . \tag{7}$$

Based on Eq.(7), $sum(H_i)$ can be performed very efficiently using one addition and two subtractions. If we consider $\pi(c,\rho)$ as a version of H_i, $\pi(c,\rho)$ can be efficiently obtained with the time complexity $O(1)$. Thus, the time for obtaining $\pi(c,\rho)$ is independent of ρ.

4.1.3 Ring Structure

To more accurately match a pattern, a ring structure with layers will be proposed here for dealing with its inner localities. With different ρ, $\pi(c,\rho)$ can form a ring structure for pattern matching as Fig. 3(b). It is a good rotation invariant feature for matching a pattern if it has different rotation changes. With $\pi(c,\rho)$, we can perform the coarse matching process for getting a set CM of possible matching candidates. Then, like Fig. 2, we enter the fine matching stage to get a set FM of fine matching candidates from CM using different ring structures for this filtering process.

In Eq.(5), if we permit the point c having different shifts, different ring structures can be generated. Let $l_i = m \cdot 2^{-(i+1)}$, $h_i = \{(\pm l_i, 0), (0, \pm l_i)\}$, and $g_i = \{(\pm l_i, \pm l_i)\}$ where m×m is the dimension of the searched pattern p. In addition, let C_s^i denote the set of all shifted points of c at the ith layer. C_s^i is generated from its previous layer


(a)

(b)

Fig. 5. Different types of ring structure with concentric integral sampling

C_s^{i-1} with the recursive form: $C_s^i = \{C_s^{i-1} + b \,|\, b \in h_i \cup g_i\}$. Initially, C_s^0 includes only the element *c*, i.e., $C_s^0 = \{c\}$. Clearly, C_s^1 owns eight elements and C_s^2 has 64 ones. At the *i*th layer, for all elements c_j^i in C_s^i, its corresponding ring structure $\pi_i(e)$ is generated as follows

$$\pi\left(c_j^i\right) = \left(k\left(c_j^i, l_{i-1}\right), k\left(c_j^i, l_{i-1}+1\right), k\left(c_j^i, l_i\right)\right). \tag{8}$$

The eight ring structures of $\pi_1(e)$ are listed in Fig. 5 (a). If two adjacent points in C_s^i are integrated, another type of ring structures can be generated like Fig. 5 (b). Then, the set of ring structures can be used to filter impossible matching candidates very efficiently.

4.1.4 Cascade Boosting Structure

Actually, each concentric feature $\pi\left(c_j^i, k\right)$ extracted from its corresponding ring structure $\pi\left(c_j^i\right)$ can form a weak hypothesis h_{ijk} for filtering impossible candidates. With the set of concentric features h_{ijk}, a cascade structure can be easily constructed for finding each desired patterns in real time. Firstly, we use the rotation-invariant concentric feature $\pi(c,k)$ to quickly filter out impossible pattern candidates under the cascade boosting structure. Then, for all the remained candidates, their different rotated versions are generated and verified using the concentric features $\pi\left(c_j^i, k\right)$ at different layers for $i > 0$. Then, the optimal location of the target can be then well detected from input images.

4.2 Scene Matching

To search the desired scene, each frame is divided to 10×10 grids like Fig. 6. Then, given an input frame *I* and a scene S_i, the matching score C_i between S_i and *I* is defined as follows

$$C_i = \sum_{k=1}^{10\times 10} B_k^I , \tag{9}$$

where B_k^I is the *k*th grid of *I* and B_t^I will be 1 if a feature pattern in S_i is matched to *I* at the grid B_t^I. The matched scene is determined as the maximum matching score for all scenes.

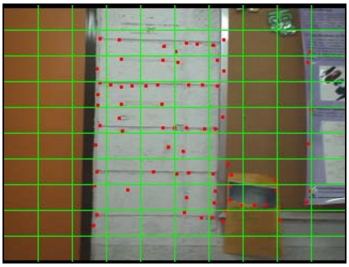


Fig. 6. Each frame is divided to different grids

4.3 Missing Object Detection

After scene matching, we present a method to detect the presence or absence of an object. Let S_M denote the best matching scene of an input frame I and F_M is the set of corners in S_M. If the camera is static, we can use a subtraction technique to detect missing objects from I. However, this technique will fail when the camera is not static. Thus, a novel spider structure is proposed for missing object detection. Firstly, we project all the corners of F_M onto I (like Fig. 7(a)) and connect two corners if they are close enough. Then, a "spider web" like Fig. 7(b) can be constructed.

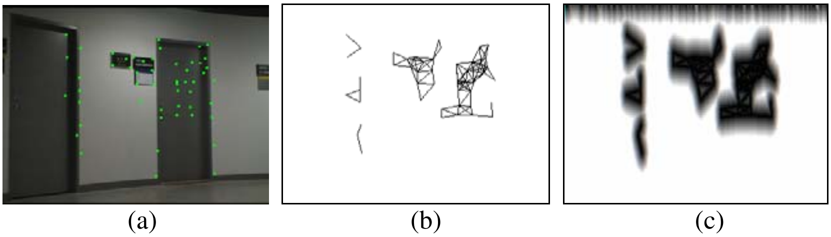

(a) (b) (c)

Fig. 7. (a) Projection result of F_M on input frame. (b) Spider web. (c) Distance map of (b).

Assume W_M and W_I are two spider webs built from S_M and I, respectively. Assume DT_{W_M} and DT_{W_I} are the distance maps of W_M and W_I, respectively. The result of distance transform of (b) is shown in Fig. 7(c). Then, each missing objects can be well detected by subtracting DT_{S_M} and DT_{S_I} . A region which contain many unmatched corners will correspond to a missing object.

5 Experimental Results

In order to analyze the performances of our proposed method, various videos captured under different weather conditions and lighting conditions were used. The dimension of video frame is 320×240. Fig. 8 shows the results of exceptional object detection

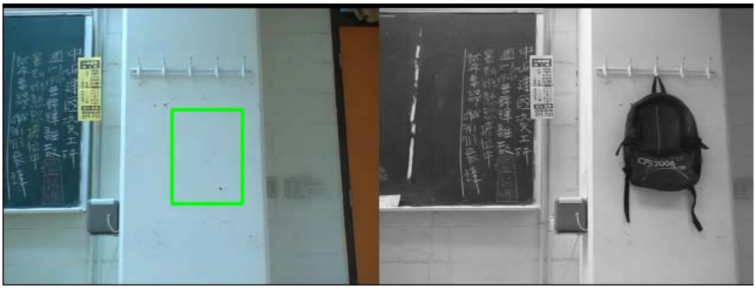


Fig. 8. Results of exceptional changes detection with complicated backgrounds

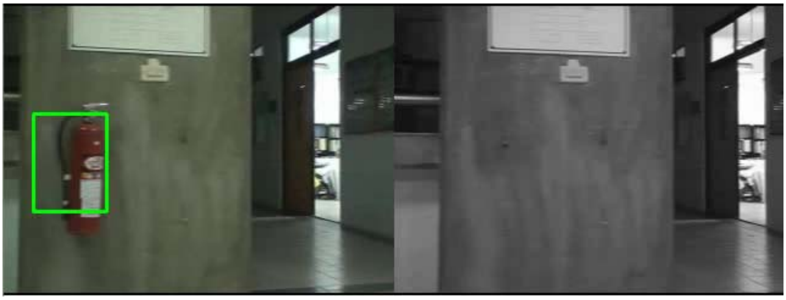


Fig. 9. Result of exceptional change detection when a corner scene was handled

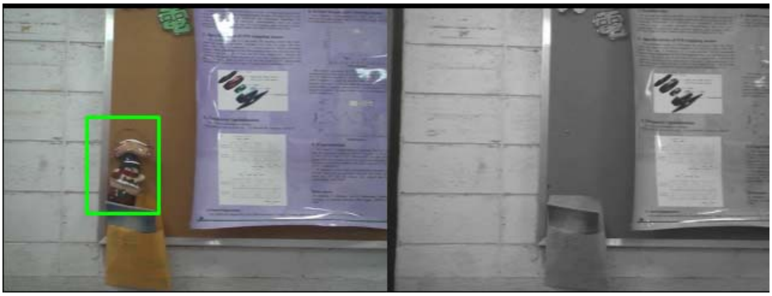


Fig. 10. Results of exceptional changes detection when a small object was handled

when complex backgrounds were handled. Fig. 9 shows the results when a corner turn was handled. It is noticed that when a scene at a corner is handled, its scene contents will change significantly. However, our method still works very well to detect a missing object. Fig. 10 shows a case when the exception object is small. All the experimental results have proved the superiorities of our proposed system in missing object detection when a moving camera is used.

6 Conclusions

We have presented a novel robust exceptional detection scheme for mobile robot surveillance system using patch matching and concentric features. In the training phase, we selected key frames by ordinal distance and tracking each patch using concentric features. The time complexity of patch match is reduced up to an order. For well detecting each abnormal object, we propose an spider web structure for foreground object detection. Even though the input frame is not well registered to the scene panoramas, our method still works well to detect each abnormal object. Experimental results have proved the superiorities of our proposed system in exceptional change detection on a mobile robot under different background and lighting conditions.

References

1. Collins, R.T., Lipton, A.J., Kanade, T.: Introduction to the special section on video surveillance. IEEE Trans. on Pattern Analysis and Machine Intelligence 22(8), 745–746 (2000)
2. Zhong, H., Shi, J., Visontai, M.: Detecting unusual activity in video. In: Proc. of IEEE Conf. on Computer Vision and Pattern Recognition, vol. 2, pp. 819–826 (2004)
3. Kim, K., et al.: Real-time Foreground-Background Segmentation using Codebook Model. Real-time Imaging 11(3), 172–185 (2005)
4. Stringa, E., Regazzoni, C.S.: Real-time video-shot detection for scene surveillance applications. IEEE Transactions on Image Processing 9(1), 69–79 (2000)
5. Foresti, G.L., Marcenaro, L., Regazzoni, C.S.: Automatic detection and indexing of video-event shots for surveillance applications. IEEE Transactions on Multimedia 4(4), 459–471 (2002)
6. Castelnovi, M., et al.: Surveillance Robotics: analyzing scenes by colors analysis and clustering. In: Proc. of IEEE International Symposium on Computational Intelligence in Robotics and Automation, vol. 1, pp. 229–234 (2003)

Video Inpainting on Digitized Old Films

Nick C. Tang[1], Hong-Yuan Mark Liao[1], Chih-Wen Su[1], Fay Huang[2]
and Timothy K. Shih[3]

[1] Institute of Information Science, Academia Sinica, Taiwan
[2] Institute of Computer Science and Information Engineering, National Ilan University, Taiwan
[3] Department of Computer Science,National Taipei University of Education,Taiwan
{nickctang,liao,lucas}@iis.sinica.edu.tw, fay@niu.edu.tw,
timothykshih@gmail.com

Abstract. Video inpainting is often used as a tool to assist user in removing objects or repairing damaged areas in a video. To deal with different kinds of video, several techniques such as object segmentation and temporal continuity maintenance are commonly adopted. In this paper, we extend the concept of exemplar-based inpainting to propose a new video inpainting algorithm which can realize object removal both in modern digital video and digitized aged films. Furthermore, a new patch searching strategy and a new patch adjustment mechanism are intorduced to maintain the temporal continuity of video and thus improve the results of video inpainting. Experiments demonstrate that the proposed algorithm can be effectively applied to different types of video.

Keywords: Image Inpainting, Video Inpainting, Object Segmentation, Motion Estimation.

1 Introduction

Since the image inpainting–related techniques have been well developed, researchers started to extend the scope from image to video [4,5,7]. An intuitive way of performing video inpainting is to treat each video frame as an independent image. However, an image inpainting process can only take care of the content of an image frame, it does not simultaneously handle the continuity issue across consecutive frames. To better design a video inpainting process, both the spatial (intra-frame) and temporal (inter-frame) continuity issues have to be considered.

In the past few years, video inpainting has become more popular due to its potential applications in daily life [4,5,7,9,10]. In [7], Patwardhan *et al.* extended the image completion concept proposed in [1] to deal with non-stationary videos. In their work, the background and foreground of a video are separated and the corresponding optical-flow mosaics are generated. After the foreground of all frames is inpainted, the remaining background holes are filled with the patches extracted directly from the adjacent frames through a texture synthesis process. In [5], Jia *et al.* proposed a two-phase approach to perform video inpainting. The two phases are a sampling phase and an alignment phase. These two phases work together can predict the motion of moving foreground and at the same time align the repaired foreground by damaged background.

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 421–430, 2009.

All the above mentioned video inpainting techniques were developed to handle modern digital videos. In recent years, transforming cultural heritages into digital format has become an important trend. However, most of the old pictures or old films that have historical value are aged. These aged images or videos, after digitization, usually turn out with very poor quality. They very often contain unstable luminance and damaged contents. In [3], Gullu *et al.* worked on video scratch detection via temporal coherency analysis. In [6], Machi and Collura conducted a research on fixing single frame defects through spatiotemporal analysis. However, their work only deal with small region defects and the most important issue - temporal continuity, was not seriously addressed.

In this paper, we proposed a new video inpainting algorithm to deal with digitized old films. Since these digitized old films were taken by old analog camcorders, their qualities are very poor. First, we apply a global intensity normalization process to adjust the average luminance level across consecutive frames. This step is to stabilize the luminance of target videos. Second, we perform an object segmentation process to identify the foreground of a video. Then, in the third step, a block-based motion estimation process is introduced to compute the motion map of each frame. This map can then be used to maintain the continuity effect across temporal axis during our video inpainting procedure. Finally, we work on the background part to complete the whole inpainting process.

2 Motion Layer Segmentation

When dealing with non-stationary video with moving object, motions of a video can be categorized into two types. The first type is local motion caused by moving objects and the second type is global motion caused by camera movement. Thus, to maintain temporal continuity precisely, the procedure of motion layer segmentation is a must. In this paper, we propose a motion layer segmentation procedure that includes three major steps. They are: (1) a normalization step that normalizes the average intensity of every frame in a video; (2) an object tracking step that segments local object motions; and (3) a camera motion estimation step that is used to compute the motion flow of each block to maintain temporal continuity.

2.1 Average Intensity Normalization of Video Frame

Most block-based motion estimation algorithms make use of the similarity comparison mechanism in their searching procedure. However, the results generated by a similarity comparison procedure are very easily influenced by unstable luminance conditions. Therefore, we try to normalize the average intensity of every frame. The proposed normalization procedure is described as follows. First, we compute the difference of the average intensity, I_{diff}, between two consecutive frames F_t and F_{t+1}, where

$$\mathrm{I_{diff}} = \frac{\sum_{\mathrm{p}} \mathrm{I(p)}}{|\mathrm{F_t}|} - \frac{\sum_{\mathrm{p'}} \mathrm{I(p')}}{|\mathrm{F_{t+1}}|} \quad , \mathrm{p} \in \mathrm{F_t} , \forall \mathrm{p'} \in \mathrm{F_{t+1}} \tag{1}$$

p and *p'* represent a pixel in frame F_t and frame F_{t+1}, respectively. *I(p)* represents the intensity of pixel *p*. $|F_t|$ and $|F_{t+1}|$ indicate the total number of pixels in frame F_t and

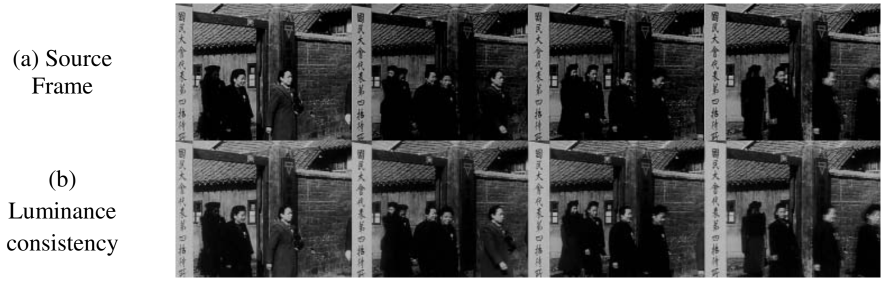


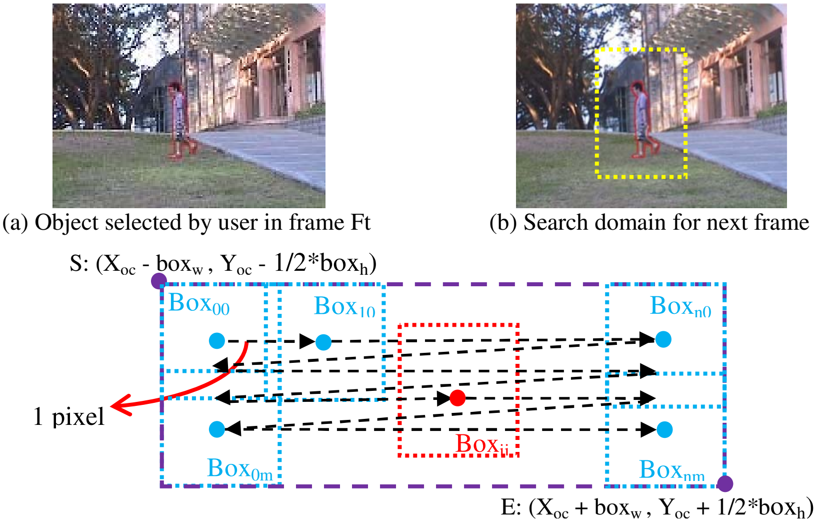
Fig. 1. Result of average intensity normalization

frame F_{t+1}, respectively. After I_{diff} is computed, we can adjust the intensity of each pixel in frame F_t to make the average intensity level between F_t and F_{t+1} closer. Fig. 1 shows the original video sequence (Fig. 1(a)) and the sequence after intensity normalization (Fig. 1(b)).

2.2 Target Object Segmentation

After the average intensity normalization process is applied across frames, the next step is to extract foreground objects from each frame. Our object segmentation process starts from object selection which is the only step that needs human intervention.

A user needs to select an object that is to be removed in the very beginning. Fig. 2(a) shows an example of object selection. After an object is chosen, we then calculate the object's center coordinates (X_{oc}, Y_{oc}). In addition, we also analyze the color information of the object. To identify the pose of the same object in the next frame,

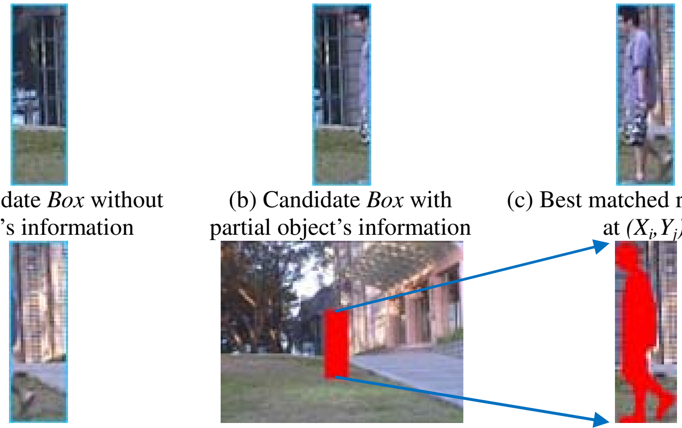

S: $(X_{oc} - box_w, Y_{oc} - 1/2*box_h)$
Box_{00}
Box_{10}
Box_{n0}
1 pixel
Box_{ij}
Box_{0m}
Box_{nm}
E: $(X_{oc} + box_w, Y_{oc} + 1/2*box_h)$

(a) Object selected by user in frame Ft

(b) Search domain for next frame

(c) Definition for searching strategy

Fig. 2. Diagram of proposed searching algorithm



(a) Candidate *Box* without object's information

(b) Candidate *Box* with partial object's information

(c) Best matched result Box_{ij} at (X_i, Y_j)

(d) Candidate *Box* with partial object's information

(e) Targeted bounding box Box_{ij} in frame F_{t+1}

(f) Segmentation of object in Box_{ij}

Fig. 3. Example of target object segmentation

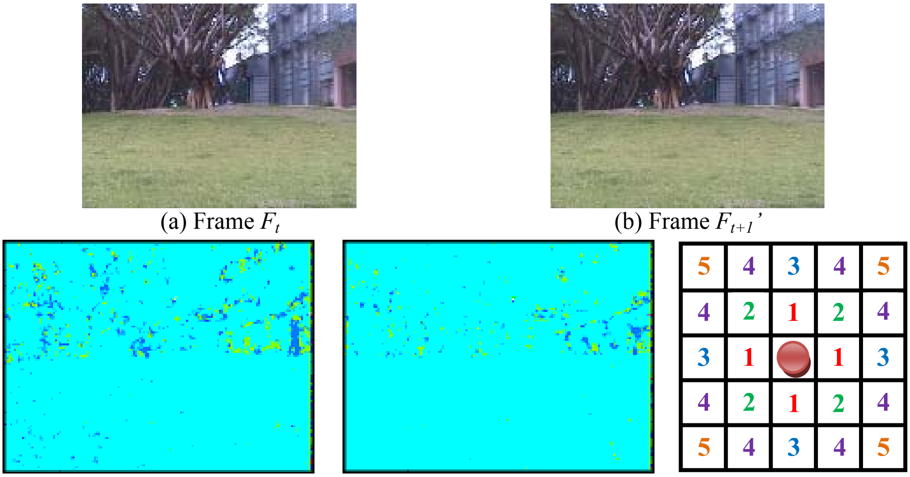
we adopt a full search strategy in a neighboring domain which is the neighborhood of the chosen object at the current frame (Fig. 2(b)). We use a bounding rectangle with width box_w and height box_h to bound the chosen target object. The top-left corner *S* of the search domain has the coordinate $(X_{oc} - box_w, Y_{oc} - 0.5 * box_h)$ and the bottom-right corner *E* has the coordinate $(X_{oc} + box_w, Y_{oc} + 0.5 * box_h)$. Then, a slide window with width box_w and height box_h is applied to perform full search within the above search domain. Fig. 2(c) shows the definition of the searching strategy. The proposed search process starts from *S* to *E* to identify the best matched result in the search domain.

Fig. 3 shows an example of our search process. Fig. 3(a) to 3(d) are examples of candidate boxes determined by the search process. Fig. 3(c) is the best matched result Box_{ij} at a specific location (X_i, Y_j). Fig. 3(e) is the corresponding bounding box of Box_{ij} in frame F_{t+1}. All blocks that contain the object's information in Box_{ij} are classified and marked according to the information of object collected in the step of object selection. Fig. 3(f) shows an example of object segmentation. Since the object is segmented, the search process will update the object's information including color and motion. These updated information will be used in the next iteration.

2.3 Construction of Motion Map

To guarantee temporal continuity, we propose to maintain a motion map for every video frame. Here, the unit used to compute motion vector is macroblock (16 x 16). This is the same as the size used in MPEG. The proposed motion map construction algorithm is as follows:

1. Use CDHS algorithm proposed in [8] to compute motion vectors from each macroblock in frame F_t (Fig. 4(a)). The color space adopted is HSI.
2. Copy all macroblocks from F_t to generate a pseudo frame F_{t+1}' (Fig. 4(b)) based on the motion vectors calculated in step 1.



(a) Frame F_t

(b) Frame F_{t+1}'

(c) Difference between F_t and F_{t+1}'

(d) After re-estimation

(e) mask of direction selection

Fig. 4. Example of motion map construction

3. Compare the differences between F_{t+1} and $F_{t+1'}$ (as shown in Fig. 4(c)). Those differences can be viewed as mis-estimated blocks.
4. For each mis-estimated block found in step 3, re-estimate the motion vectors.
5. Construct a motion map based on the motion vectors calculated from each macroblock.

In step1, we calculate the Sum of Squared Differences (SSD) between a source macroblock B_i and a target macroblock B_j based on the HSI color components, respectively. The similarity measurement function is as follows:

$$\mathrm{dis}(B_i, B_j) = H_{SSD_{(i,j)}} + S_{SSD_{(i,j)}} + I_{SSD_{(i,j)}} \tag{2}$$

B_i is distributed by nature and its distribution is based on the CDHS searching topology defined in [8].

After executing step 1, the motion vector of each block in Frame F_t is determined and these motion vectors are then used to generate a pseudo frame F_{t+1}'. Fig. 4(b) shows an example of a pseudo frame. The functionality of frame F_{t+1}' is to identify those mis-estimated macroblocks by comparing its contents with those of frame F_{t+1}. Once F_{t+1}' is generated, the motion vectors of those mis-estimated macroblocks will be re-estimated with a new initial direction, d_i, and the CDHS searching algorithm will be run again. As to the selection of d_i, it is determined by those correctly estimated surrounding blocks. Fig. 4(e) shows a mask used to perform direction selection. One can search the surrounding blocks (ordered by digits from 1 to 5) to identify a valid motion vector and then use it in the searching process. However, if all surrounding blocks are mis-estimated, one can randomly choose a direction to which is associated with a second higher value to proceed. Fig. 4(d) shows the result of re-estimation which indicates that the errors are significantly reduced. Since there is no mis-estimated blocks in F_t, the motion map of frame F_t can be generated by collecting the motion vectors of all blocks. The motion map of a frame reflects the motion direction of every block in that frame and these information can be used to maintain the temporal continuity.

3 Video Inpainting by Maintaining Temporal Continuity

Temporal continuity is very important in the process of video inpainting. A video inpainting algorithm that does not take temporal continuity into account will not generate visually pleasing inpainting results. Our video inpainting algorithm mainly consists of two parts. They are a priority map computation step and a patch search and adjustment step. We introduce new strategies in the patch searching and adjustment process. These new strategies can improve the inpainting results while at the same time maintaining the temporal continuity of an inpainted video.

3.1 Priority Map Computation

Priority map is basically the concept proposed by Criminis *et al.* [1]. We modify the data term of their scheme to make the computation faster and the maintenance of the information structure easier. As shown in Fig. 5, let $\boldsymbol{I}$ be a frame of a video which includes a source area $\boldsymbol{\Phi}$ and a target area $\boldsymbol{\Omega}$ to be inpainted. Hence, $\boldsymbol{I} = \Phi \cup \Omega$. The notation $\delta\Omega$ is the front contour on Ω and Ψ_p is an arbitrary patch centered at pixel $p \in \delta\Omega$ (right hand side of Fig. 5).

The difference between our approach and [1] is that we modified the definition of their data term based on the observation of continuous structure derivation. We use a mean shift region segmentation algorithm [2] to segment the original frame into several regions. The mean shift procedure takes $\boldsymbol{I}$ as input, and produces $\boldsymbol{I'}$ as the result of segmentation. An edge detection algorithm is then used to convert $\boldsymbol{I'}$ to a binary image $\boldsymbol{BI}$, which results in an edge map. The reason why we make use of the edge information in the data term is as follows. Let $\Phi\varepsilon \subset \boldsymbol{BI}$ be the area corresponding to $\Phi \subset \boldsymbol{I}$. $\Phi\varepsilon$ is the edge map of the source area Φ. After an edge map $\boldsymbol{BI}$ is obtained, our video inpainting procedure will proceed to compute the priority of each patch to decide on the order of inpainting. The computation of priority is based on the rule set in [1] which mainly includes two terms: a confidence term $C(p)$ and a data term $D(p)$. The explicit expression of a priority count is as follows:

$$P(p) = C(p)*D(p). \tag{6}$$

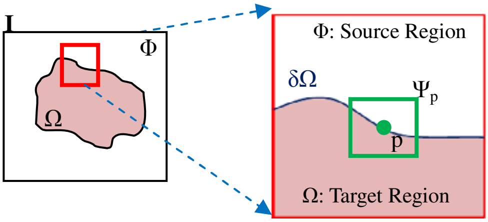


Fig. 5. Notation used in our video inpainting algorithm

The function of the priority term is to determine the inpainting order of each constituent patch. The confidence term is used to check the degree of damage of a target patch. Before we compute the confidence term, the initial confidence value of each pixel in ***I*** is assigned at first. That is,

$$\mathrm{C(p)} = \begin{cases} 1.0 & \text{if and only if p} \in \Phi \\ 0.0 & \text{if and only if p} \in \Omega \end{cases} \tag{7}$$

As indicated in Fig. 5, Ψ_p is a patch centered at pixel *p*. The formal definition of *C(p)* of an arbitrary patch Ψ_p is as follows:

$$\forall\ \mathrm{p} \in \delta\Omega, \mathrm{q} \in \Psi\mathrm{p}, \qquad \mathrm{C(p)} = \frac{(\sum_{\mathrm{q} \in (\Psi\mathrm{p} \cap \Phi)} \mathrm{C(q)}\ )}{|\Psi\mathrm{p}|} \tag{8}$$

where $|\Psi_p|$ is the area of Ψ_p. The confidence term *C(p)* in Eq. (8) is used to compute the percentage of undamaged pixels of a patch $\Psi_{p.}$ A patch with higher confidence value means its damaged area is small and its priority in a video inpainting process is higher. Here, a confidence term can be used to provide a rough guide for deciding on where to start an inpainting process. Usually, there would be a number of patches that receive high confidence simultaneously, but we have to determine among them which one is the best. Under these circumstances, a data term is introduced to work with the confidence term to solve the above mentioned problem.

As to the data term, it is used to judge how important the content of a target patch is. We propose to compute the percentage of edge amount and the degree of complexity of constituent colors in a patch to decide on whether the patch is important or not. The formal definition of our data term is as follows:

$$\forall\ \mathrm{p} \in \delta\Omega, \mathrm{q} \in \Psi\mathrm{p},\ \ \mathrm{D(p)} = \frac{\max\,(1, \sum_{\mathrm{q} \in (\Psi\mathrm{p}\ \cap\Phi\varepsilon))} c) * var(\Psi\mathrm{p})}{|\Psi\mathrm{p}|} \tag{9}$$

$$\mathrm{var}(\Psi\mathrm{p}) = \frac{\sum_{\forall\mathrm{i}} \sum_{\forall\mathrm{j}} \mathrm{c_{ij}}}{\mathrm{i} * \mathrm{j}} + \frac{\sum_{\forall\mathrm{i}} \sum_{\forall\mathrm{j}} \mathrm{l_{ij}}}{\mathrm{i} * \mathrm{j}} \tag{10}$$

where the *max* function is used to assure a non-zero summation of pixel count in the edge map and *var(*Ψ_p*)* is used to compute the degree of color variation in patch Ψ_p.

3.2 Patch Searching and Further Adjustment

After the priority map is determined, the next step is to select a best-matched patch based on the value of the priority computed. Our search strategy is that if a patch receives a highest priority value, it will have better chance to find a patch with the best quality. This is because in the priority map determination process we have put the degree of damage, the relation with surrounding patches, and the degree of complexity into consideration. These embedded characteristics can ensure the quality of a search process. Let $\Psi_{p^\wedge}$ be a patch that receives the highest priority, where

$$\mathrm{p}^\wedge = \mathit{argmax}(\mathrm{P(p)}, \mathrm{p} \in \delta\Omega). \tag{11}$$

An arbitrary patch template, $\Gamma_{p^\wedge}$, of $\Psi_{p^\wedge}$ can be expressed as follows:

$$\Gamma_{\mathrm{p}^\wedge} = \cup_{\pm k\Psi\mathrm{p}^\wedge} (\Psi_{\mathrm{p}^\wedge} \cap \Phi) \neq \phi, \text{ where } \phi \text{ is an empty set.} \tag{12}$$

Here, $\pm k\Psi_{p^\wedge}$ represents a domain that covers the patch $\Psi_{p^\wedge}$ plus its surrounding pixels. These neighboring pixels are within a distance of *k*. The value *k* is set as one half width of a patch. In all of our experiments, the size of patch is 3*3. To ensure no "empty patch" is mis-used, we also cover part of the intersection region between Φ and Ω. After defining the patch template, one can start the searching process.

Since in the previous stages we have built the motion map for the background, the search of an appropriate patch can be done through the assistance of motion vectors. This searching process will start from neighboring blocks and then those blocks that are farther. If the previous step cannot identify useful blocks, then the algorithm will trigger an intra-frame search to locate a most similar patch to do inpainting. Let $\Gamma_{q^\wedge}$ be the best matched patch template against all candidate patch templates, that is

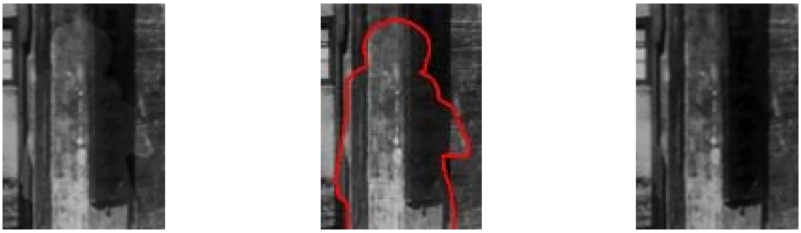
$$\forall q \in P_{s,}\ \Gamma_{q^\wedge} = min_{\Gamma q \in \Phi q^{r}}\, d(\Gamma_{p^\wedge}, \Gamma_q), \tag{13}$$

where $\Phi_q^{\ r} \subset \Phi$ represents a region of source image centered at q and within the distance of r pixels. $d(\Gamma_{p^\wedge}, \Gamma_q)$ is a distance function that can measure the difference between $\Gamma_{p^\wedge}$ and $\Gamma_{q.}$ The formal definition of this distance is as follows:

$$d(\Gamma_{p^\wedge}, \Gamma_q) = SSD\,(\Gamma_{p^\wedge}, \Gamma_q) * max(1, (\textstyle\sum_{q \in (\Gamma q\ \cap \Phi \varepsilon)} c)). \tag{14}$$

SSD here represents the sum of squared distance. The distance function defined in Eq. (14) considers two factors, i.e., the SSD and the number of useful pixels in the patch template. We want to find a best patch $\Psi_{q^\wedge}$, where $\Psi_{q^\wedge} \subset \Gamma_{q^\wedge}$. Hence, the inpainting algorithm copies $\Psi_{q^\wedge}$ to $\Psi_{p^\wedge}$, $\forall p \in \Psi_{p^\wedge} \cap \Omega$ (i.e., only copy pixels to the empty positions in patch $\Psi_{p^\wedge}$, without changing the area already inpainted). The constant, c=1, represents the weight of a useful pixel.

After a target patch is identified from a referenced frame, a patch adjustment process is proposed to repair the artifacts caused by unstable illumination conditions. The patch adjustment process proceeds as follows: (1) calculate the difference between the intensities of the surrounding pixels of $\Psi_{q^\wedge}$ and $\Psi_{p^\wedge}$, respectively; (2) regulate the intensity of $\Psi_{q^\wedge}$ by applying the gamma correction technique on a target patch. An example of patch adjustment is shown in Fig. 6. Fig. 6(a) shows the artifact caused by pasting a path to a damaged area directly. The red silhouette of Fig. 6(b) indicates the mask which is the surround pixels of $\Psi_{p^\wedge}$ and used to calculate the difference of intensity between $\Psi_{q^\wedge}$ and $\Psi_{p^\wedge}$. Fig. 6(c) shows that the intensity of the targeted patch is adjusted and pasted in corresponding region.



(a) without patch adjustment (b) Boundary of target patch (c) with patch adjustment

Fig. 6. Comparisons of video inpainting procedure with or without patch adjustment

4 Experimental Results

We used several types of video examples for experiments, including home videos (Fig. 7(a)) and aged films (Fig. 7 (c) and Fig. 7(e)) with different camera motions such as panning and hand shaking. The corresponding results after applying our video inpainting algorithm are shown in Fig. 7. Figs. 7(a), (c), (e) is the source sequences and Figs. 7(b), (d), (f) are the corresponding results after applying our video inpainting algorithm.

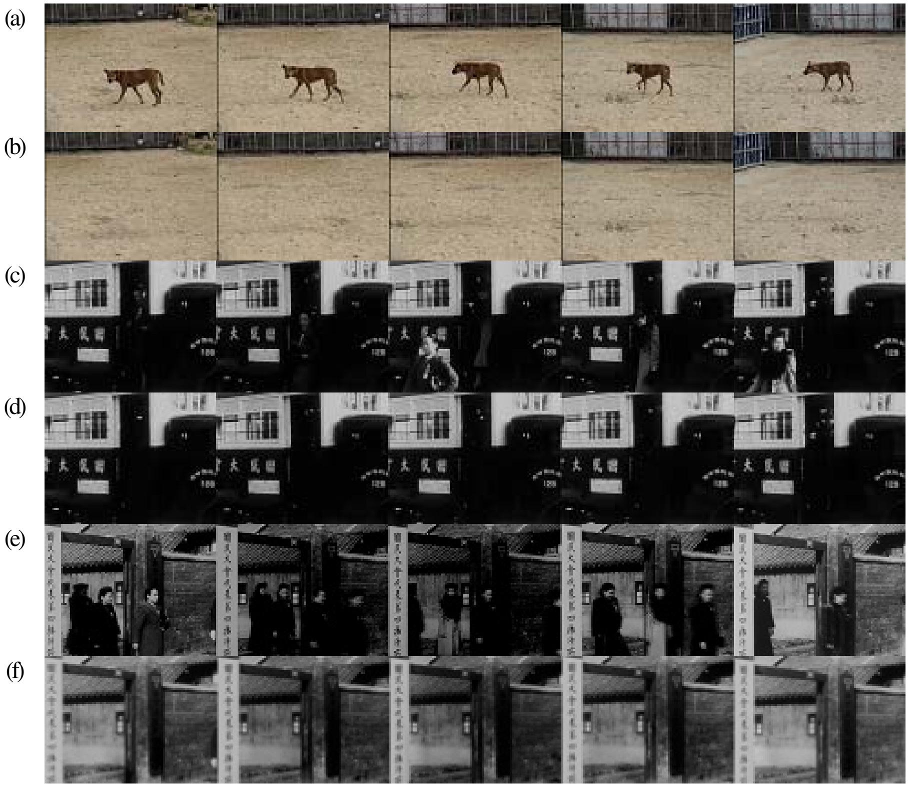


Fig. 7. Video inpainting in different kinds of video

5 Conclusion

In this paper, we proposed an algorithm which can realize video inpainting both in modern digital videos and digitized aged films. The experimental results show our algorithm can produce visually pleasant inpainting result. The proposed video inpainting procedure, based on the analysis of temporal continuities of video, is able to deal with different camera motions. Although video with really bad quality cannot be inpainted with all visual defects removed, we will try some technologies such as poisson image editing in our patch adjustment procedure to improve our result in the future.

References

[1] Criminisi, A., Perez, P., Toyama, K.: Region Filling and Object Removal by Exemplar-Based Image Inpainting. IEEE Trans. Image Processing 13, 1200–1212 (2004)

[2] Comaniciu, D., Meer, P.: Mean Shift: A Robust Approach toward Feature Space Analysis. IEEE Trans. on Pattern Analysis and Machine Intelligence 24(5) (May 2002)

[3] Gullu, K.M., Urhan, O., Erturk, S.: Scratch Detection via Temporal Coherency Analysis. In: Proc. of the 2006 IEEE International Symposium Circuits and Systems (2006)

[4] Jia, J., Wu, T.-P., Tai, Y.-W., Tang, C.-K.: Video Repairing: Inference of Foreground and Background under Severe Occlusion. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June-July 2004, pp. I-364–I-371 (2004)

[5] Jia, J., Tai, Y.W., Wu, T.P., Tang, C.K.: Video Repairing Under Variable Illumination Using Cyclic Motions. IEEE Trans. on Pattern Analysis and Machine Intelligence 28(5) (2006)

[6] Machi, A., Collura, F.: Accurate spatio-temporal restoration of compact single frame defects in aged motion pictures. In: Proc. of the 12th International Conference on Image Analysis and Processing, pp. 454–459 (2003)

[7] Patwardhan, K.A., Sapiro, G., Bertalmío, M.: Video Inpainting Under Constrained Camera Motion. IEEE Trans. on Image Processing (February 2007)

[8] Cheung, C.H., Po, L.M.: Novel cross-diamond-hexagonal search algorithms for fast block motion estimation. IEEE Trans. on Multimedia 7(1), 16–22 (2005)

[9] Wexler, Y., Shechtman, E., Irani, M.: Space-Time Completion of Video. IEEE Trans. on Pattern Analysis and Machine Intelligence 29(3), 463–476 (2007)

[10] Zhang, Y., Xiao, J., Shah, M.: Motion Layer Based Object Removal in Videos. In: The Seventh IEEE Workshops on Application of Computer Vision, pp. 516–521 (2005)

Pedestrian Identification with Distance Transform and Hierarchical Search Tree[*]

Daw-Tung Lin and Li-Wei Liu

Department of Computer Science and Information Engineering
National Taipei University
151, University Rd., San-Shia, Taipei, 237 Taiwan
dalton@mail.ntpu.edu.tw

Abstract. This work develops a novel and robust hierarchical search tree matching algorithm, in which the Distance Transform based pedestrian silhouette template database is constructed for efficient pedestrian identification. The proposed algorithm was implemented and its performance assessed. The proposed method achieved an accuracy of 89% true positive, 92% true negative and low false positive 8% rates when matching 1069 pedestrian objects and 568 non-pedestrian objects. The contributions of this work are twofold. First, a novel pedestrian silhouette database is presented based on the Chamfer Distance Transform. Second, the proposed hierarchical search tree matching strategy utilizing Fuzzy C-means clustering method can be adopted for mapping and locating pedestrian objects with robustness and efficiency.

Keywords: Video surveillance, pedestrian identification, distance transform, pedestrian silhouette, hierarchical search tree.

1 Introduction

Pedestrian identification and tracking is an important and yet challenging task for video surveillance. Various surveillance applications requires the detection of pedestrian to ensure safety or traffic management in daily activities of interest such as in transportation stations, public buildings, business market places, etc.. Much progress has been made in the detection and tracking of people in the computer vision research community. Wren *et al.* propose a real-time system "Pfinder" for tracking people and interpreting their behavior using a multiclass statistical model [1]. Another typical model is "W^4" proposed by Haritaoglu *et al.* [2]. W^4 is a real time visual surveillance system for detecting and tracking multiple people in an outdoor environment. Stauffer and Grimson develop a visual monitoring system that passively observes moving objects and learns patterns of activity [3]. Sidla *et al.* presents a vision based pedestrian detection

[*] This work was supported in part by the National Science Council, Taiwan, R.O.C. grants NSC96-2221-E-305-008-MY3 and Ministry of Economics grant No. 97EC17A02S1032 Construction of Vision-Based Intelligent Environment (II). Mr. Shun-Chun Lee is also commended for his work in the preliminary study.

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 431–438, 2009.

and tracking system for crowded situations [4]. Masoud and Papanikolopoulos presents a real-time system for pedestrian tracking in sequences of grayscale images acquired by a stationary camera for pedestrian control scheme at intersections [5]. There are a number of challenging issues worth further exploration such as camera viewpoints, distance between human and camera, non-rigid human motion, occlusion, recognition speed, etc. In this paper, we attempt to emphasize on the real-time pedestrian identification and present a silhouette-based method to identify humans and this method is robust enough to detect objects of various shapes. Inspired by [6], matching involves a hierarchical search tree approach over a distance-transformed images database hierarchy structure. Thus, the proposed method is able to generate an efficient representation from off-line template examples, matching proceeds on-line based on a hierarchical database which is established in advance.

The remainder of this paper is organized as follows. Section 2 reviews the Distance-Transform (DT) approach and presents the construction of pedestrian silhouette database. Section 3 illustrates the object matching algorithm. Section 4 proposes DT-based fast matching algorithm with hierarchical search tree for pedestrian identification. Section 5 analyzes the experimental results. Conclusions are finally drawn in Section 6.

2 Distance Transform and Pedestrian Silhouette Database Construction

The proposed system initially builds a statistical background model by combining the Gaussian Mixture Model [7] and the approach proposed by KaewTraKulPong *et al.* [8] which leads in better performance in terms of learning speed and accuracy [9]. A pixel is considered as a background, if matched with one of the Gaussian distributions. Otherwise, the pixel is marked as a part of a moving object (foreground). Then, the connected component operation is perform. To make up the fragment effect of segmentation in foreground region, we employ the opening and closing of morphological operations. Furthermore, to reduce the susceptibility problems arisen from shadow influence, we employ the HSV color space to distinguish shadow pixels from the ordinary foreground objects. Afterwards, we extract the silhouettes of moving objects from the resultant foreground objects and select the silhouettes of moving pedestrian manually to construct a pedestrian silhouette template database. To cluster these template

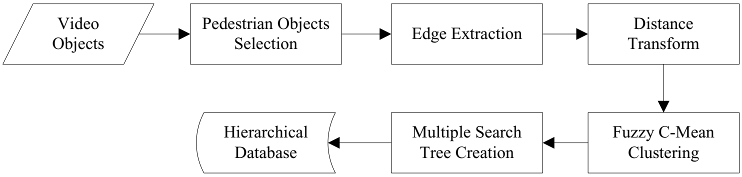


Fig. 1. The flow chart of constructing pedestrian silhouette database

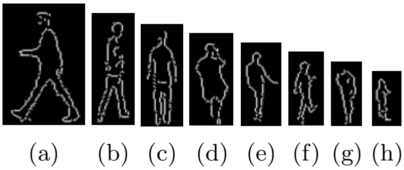


Fig. 2. Examples of different heights of pedestrian silhouette template images: (a) 92 pixels, (b) 84 pixels, (c) 76 pixels, (d) 68 pixels, (e) 60 pixels, (f) 52 pixels, (g) 44 pixels and (h) 36 pixels

data, we extract the edge image from the selected silhouettes by using Canny Edge detection and then compute the distance transform for each Canny Edge image of object silhouettes that are considered as pedestrian. Fig. 1 illustrates the schematic flow chart of the pedestrian silhouette database construction procedure.

The edge image S_n^i of the selected silhouettes is given as

$$S_n^i = \left\{(x, y)|E_n(x, y) = 1 \text{ and } (x, y) \subseteq Obj_n^i, \right\}, \qquad (1)$$

where S_n^i denotes the edge set of the i^{th} foreground object Obj_n^i in the n^{th} frame extracted from the above mentioned background substraction and shadow justification procedure. $E_n(x, y)$ represents the edge image calculated via Canny Edge Detection operation on the foreground object image. To identify pedestrians in a video sequence, we need to collect a lot of template images of pedestrians with different sizes, walking directions and various shapes. In this paper, we classify all template images into 24 categories based on different image heights. Fig. 2 shows eight examples of different heights.

The next step is to estimate the similarity between two images. Typically, Euclidean distance is the traditional method of evaluating the similarity of two images. However, the high computational complexity makes it difficult to implement in real-time applications. Thus, we apply the 3-4 Chamfer Distance Transform (3-4 DT) due to its simplicity and robustness [10]. The DT equation is shown in below.

$$DT(x + dx, y + dy) = \min \{ra + sb\}, \qquad (2)$$

where r denotes the horizontal or vertical moving steps, s is the diagonal moving steps and a, b are the weights of two different steps (i.e., $a = 3$, $b = 4$ in 3-4 DT). Thus, the value of $(x+dx, y+dy)$ of 3-4 DT is to find the shortest distance from (x, y) to $(x + dx, y + dy)$.

Fig. 3 illustrates two examples of DT images. Fig. 3(a) and (d) are the original object images, Fig. 3(b) and (e) show the edge images after Canny Edge Detection, and Fig. 3(c) and (f) present the resultant Distance Transform of the corresponding edge images. The distances in the DT images are intensity-coded, i.e., the gray value of (x, y) represents the distance to the nearest edge point where lighter color denotes larger distance and darker color indicates shorter distance.

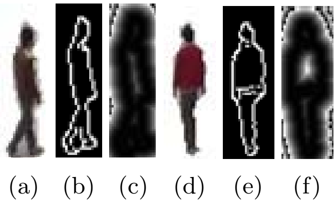


Fig. 3. Examples of edge detection and distance transform. (a) and (d) are the original object images, Fig. 3(b) and (e) show the edge images after Canny Edge Detection, and Fig. 3(c) and (f) present the resultant Distance Transform of the corresponding edge images.

3 Object Matching Algorithm

Let S be the source edge image, E be the target edge image. The DT image DT_S, DT_E of S and E can be computed by using Equation (2). Therefore, the chamfer distance of these two images is defined as

$$EDist_{S,E}(dx, dy) = \frac{1}{|S|} \sum_{(x,y) \in S} DT_E(x + dx, y + dy), \tag{3}$$

where $|S|$ denotes the number of pixels in S and (dx, dy) denotes the shift coordinate when the source edge image S map to the target edge image E by adding the coordination shift value (dx, dy). According to the definition of DT image, the lower distance is, the better match between S and E at location $(x + dx, y + dy)$ is. Thus, our goal is to find the minimum chamfer distance by modifying shift coordinate (dx, dy) repeatedly, (i.e., $\min EDist_{S,E}(dx, dy)$). How to find the minimum $EDist_{S,E}(dx, dy)$ in real-time application can be considered as the motion estimation problem. One of the simplest matching algorithm is the full search (FS) algorithm which provides an optimal solution by searching all candidates within the target edge image E. There are a number of matching algorithms can reduce the time complexity of FS algorithm. We apply the three step search (TSS) algorithm in which the matching criterion of TSS is replaced by the minimum chamfer distance $\min EDist_{S,E}(dx, dy)$. Fig. 5 depicts the block diagram of template image matching algorithm.

However, the experimental result shows that when the number of edges of S is far less than that of E, the chamfer distance $EDist_{S,E}(dx, dy)$ will also be small and this observation of minimum distance could lead in a flaw. Therefore we need to modify Equation (3) to overcome this drawback as follows.

$$EDist'_{S,E}(dx, dy) = \frac{1 + EDist_{S,E}(dx, dy)}{1 - (||S| - |E||)/(|S| + |E|)}. \tag{4}$$

When the difference between S and E is very large, the term $(||S| - |E||)/(|S| + |E|)$ in Equation (4) will approximate to 1 so that the denominator of Equation (4) becomes a small number and $EDist'_{S,E}(dx, dy)$ will become more larger

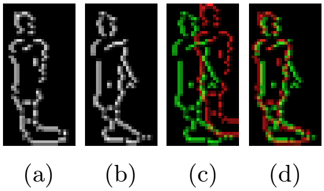


Fig. 4. Example of minimum distance searching: (a) target edge image E, (b) source edge image S, (c) $Edist_{S,E}(6,1) = 30.7$, and (d) $Edist_{S,E}(-2,1) = 4.6$

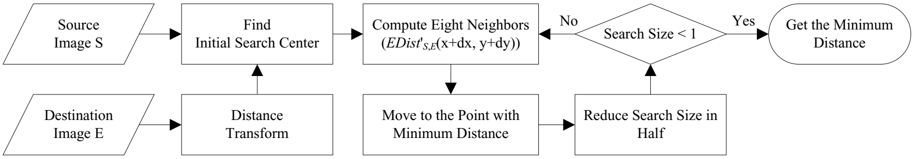

Source Image S
Find Initial Search Center
Compute Eight Neighbors ($EDist'_{S,E}$(x+dx, y+dy))
No
Search Size < 1
Yes
Get the Minimum Distance
Destination Image E
Distance Transform
Move to the Point with Minimum Distance
Reduce Search Size in Half

Fig. 5. The block diagram of template image matching algorithm

than the original distance. Only when $|S|$ and $|E|$ are similar, the measure $EDist'_{S,E}(dx, dy)$ will be a smaller value. Finally, the modified chamfer distance $EDist'_{S,E}(dx, dy)$ will be more appropriate.

4 Fast Pedestrian Identification with Hierarchical Search Tree

Matching a foreground object one by one with N template images is inefficient. We can construct a matching hierarchy with coarse-to-fine search mechanism. The idea is that each search level is coarsely divided with larger chamfer distance. Thus, we could speed up the matching process by grouping similar template images together and representing them by an instance image, rather than matching the individual template images one by one. The grouping algorithm is performed in various levels and results in a hierarchy search tree. The clustering is achieved iteratively by utilizing a top-down approach and applying a "Fuzzy C-means"-like algorithm at each level of the hierarchical structure. The distance measure of Fuzzy C-means is replaced with the chamfer distance $EDist'_{S,E}(dx, dy)$ and the cluster center is chosen as

$$C_i = \min_{S \in G_i} \sum_{E \in G_i} EDist'_{S,E}(dx, dy), \tag{5}$$

where C_i denotes the center of ith cluster and G_i is the set of ith cluster.

To identify each object Obj_n^i extracted from the current video frame, the system will read the pedestrian silhouette database given by Section 2 and create the hierarchical search tree for recognition. Then we compute the Canny edge image $SObj_n^i$ of the ith object by using Equation (1) and the distance transform

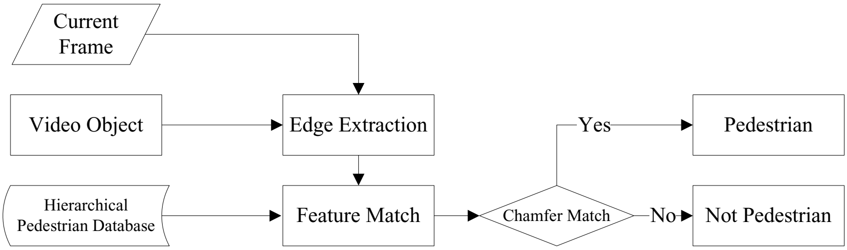


Fig. 6. The flow chart of the proposed pedestrain identification procedure

DT_S. Hence, for each video object, we will traverse in the pedestrian silhouette database and use $SObj_n^i$ and DT_S to compare the node with the chamfer distance between the node and the video object. A video object is considered match if the distance measure $EDist'_{SVO,E}(dx, dy)$ is below a pre-defined threshold θ: $EDist'_{S,E}(dx, dy) \leq \theta$. Fig. 6 shows the flowchart of pedestrian identification algorithm.

5 Experimental Results

To evaluate the performance of the proposed pedestrian identification system, we first constructed a human silhouette database using a 20 minutes video clip containing about 18000 frames. All moving objects were extracted using the procedure illustrated in Section 2, then the human objects were selected manually and form the pedestrian silhouette template database. We finally compiled a database of about 20000 pedestrian silhouette, divided and normalized them to

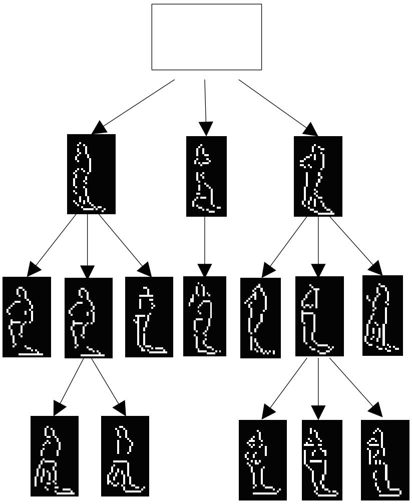
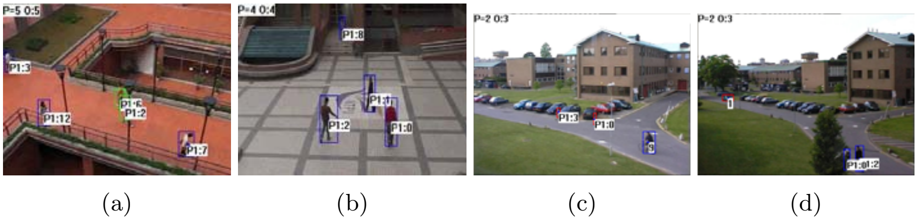


Fig. 7. A hierarchical search tree for pedestrian identification (partial view)

Table 1. Experimental results of pedestrian identification

1000 Frames	True	False
Positive	996 (89%)	122 (8%)
Negative	446 (92%)	73 (11%)



(a) (b) (c) (d)

Fig. 8. Some detection results of pedestrian identification: (a) and (b) show the results of pedestrians; (c) and (d) show the results of non-pedestrian objects

24 scales according to their height ranging from 20 to 112 pixels with step size 4 pixels. Thus, we obtained a hierarchical search tree. Fig. 7 shows a partial view of the hierarchical search tree. We can observe from Fig. 7 that the objects possess high similarity in the leaf level.

The experiments were conducted on another 1000 frames video sequence which was distinct from the template sequence. The test video clips contains 1069 pedestrian objects and 568 non-pedestrian objects. The proposed system results in a detection rate of about 89% true positive, 92% true negative and low false positive 8% rate as shown in Table 5. Fig. 8 demonstrates a few detection results on several test video clips in which label "P1" denotes the object is identified as a pedestrian. The number follows label "P1" is the object ID. Notably, Fig. 8(a) shows that our identification method can find pedestrian even with occlusion occurs (indicated in green bounding boxes). Figures 8(c) and (d) show the identification of non-pedestrian with people riding the bicycle (object ID:9) and moving vehicle (object ID:1), respectively.

6 Conclusion

This work has two principal contributions. First, an off-line pedestrian silhouette database is constructed based on 3-4 Chamfer Distance Transform. The second contribution is that this work presents a novel hierarchical search tree matching strategy utilizing Fuzzy C-means clustering method. Therefore, the proposed scheme can be adopted for various size of pedestrian matching in robust and efficient manner. The proposed method achieved an accuracy of 89% true positive, 92% true negative and low false positive 8% rates when matching 1069 pedestrian objects and 568 non-pedestrian objects. The proposed method can be easily extended to people tracking and counting applications in surveillance.

To improve the system performance, the proposed system still has some issues to be further explored, e.g. the human shape variation and object occlusion. These problems pose a strong challenge for fully automated pedestrian identification and tracking.

References

1. Wren, C.R., Azarbayejani, A., Darrell, T., Pentland, A.P.: Pfinder: Real-time tracking of the human body. IEEE Transactions on Pattern Analysis and Machine Intelligence 19(7), 780–785 (1997)
2. Haritaoglu, I., Harwood, D., Davis, L.S., Center, I.B.M.A.R., San Jose, C.A.: W4: real-time surveillance of people and theiractivities. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(8), 809–830 (2000)
3. Stauffer, C., Grimson, W.E.L.: Learning patterns of activity using real-time tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(8), 747–757 (2000)
4. Sidla, O., Lypetskyy, Y., Brandle, N., Seer, S.: Pedestrian detection and tracking for counting applications in crowded situations. In: Proceedings of the IEEE International Conference on Video and Signal Based Surveillance, pp. 70–75. IEEE Computer Society Press, Washington (2006)
5. Masoud, O., Papanikolopoulos, N.P.: A novel method for tracking and counting pedestrians in real-timeusing a single camera. IEEE Transactions on Vehicular Technology 50(5), 1267–1278 (2001)
6. Gavrila, D.M., Giebel, J., Perception, M., Res, D.C., Ulm, G.: Shape-based pedestrian detection and tracking. In: IEEE Intelligent Vehicle Symposium, vol. 1 (2002)
7. Stauffer, C., Grimson, W.E.L.: Adaptive background mixture models for real-time tracking. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 246–252 (1999)
8. KaewTraKulPong, P., Bowden, R.: An improved adaptive background mixture model for real-time tracking with shadow detection. In: Proc. 2nd European Workshop on Advanced Video Based Surveillance Systems, AVBS 2001 (2001)
9. Lin, D.-T., Lee, H.-C.: Intelligent surveillance system for halt detection and people counting. Journal of Information Technology and Applications 2(3), 133–142 (2008)
10. Borgefors, G.: Hierarchical chamfer matching: a parametric edge matching algorithm. IEEE Transactions on Pattern Analysis and Machine Intelligence 10(6), 849–865 (1988)

An Enhanced Layer Embedded Pedestrian Detector

Duan-Yu Chen

Department of Electrical Engineering, Yuan-Ze University, Taiwan
dychen@saturn.yzu.edu.tw

Abstract. We propose a method that can detect pedestrians in a single image based on the combination of Adaboost learning with a local histogram features. Besides, instead of using the raw image for further processing, we introduce a layer enhanced by orientation filters which are superimposed to the original image. Experimental results obtained using the INRIA dataset show the superior performance of our method and thus demonstrate its robustness with the novel enhanced layer embedded pedestrian detector.

Keywords: Pedestrian Detection, Orientation Filter.

1 Introduction

Automatic pedestrian detection has attracted much attention in recent years, especially in the field of surveillance and content-based image/video retrieval. Though much progress has been made [1][4][6-7], recognizing pedestrians with high accuracy remains a challenging, unsolved problem. Significant difficulties of pedestrian detection in the surveillance setting include the facts that targets often have complex shapes due to their non-rigidness and can be of low-resolution because of the nature of a video camcorder. Matters are complicated further because of the sheer range of objects that can be observed within an environment. In general, pedestrian detection involves two vital components: feature extraction and classifier design. Feature extraction is the process of deriving a set of features from the original target image domain information. Effective features for classification should minimize within-class variations of targets while simultaneously maximizing between-class variations. If inadequate features are used, even the best classifier could fail to achieve accurate recognition. In [4], Viola and Jones proposed a boosted cascade for fast face detection. This kind of cascaded classifiers has been further extended in many other object detection problems, such as pedestrian detection [9], in which rectangles features are employed to construct the weak learners of the AdaBoost classifier for each stage of the cascade. Dalal and Triggs [7] proposed a people detection method for single images. The gradient-based features, histograms of oriented gradients (HOG), are developed to describe local gradient-orientation structure. In [6], Chen and Chen proposed a method that employed both intensity-based rectangle features and gradient-based 1D features in the feature pool for weak-learner selection. The Real AdaBoost algorithm was used to select critical features from a combined feature set. Instead of using the standard boosted cascade, they employed both the stage-wise classification information and the inter-stage cross-reference information.

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 439–446, 2009.

In the previous work, they focused on the feature extraction and/or the design of training strategies to achieve higher performance of pedestrian detection. However, the way to obtain better image resources from image enhancement is not addressed in the pedestrian detection problem. Since the success of gradient-based approach depends on the contrast between edges and non-edges regions in an image. The performance of gradient-based approaches can greatly benefit by improving the image contrast. In our previous work [11], it creates directional maps by determining the dominant direction of motion in each local spatiotemporal region, and then uses the maps in real-time surveillance settings to detect pre-defined activities. Therefore, in this work, in order to strengthen the desired dominant orientating edges for boosting pedestrian detection, we introducing a novel pedestrian detector that is with an enhanced layer embedded by employing orientation filters.

The remainder of this paper is organized as follows. In Section 2, an enhanced layer obtained by using orientation filters is shown. In Section 3, the AdaBoost learning algorithm is introduced. In Section 4, a set of features based on HOG is discussed. The experimental results are demonstrated in Section 5. In Section 6, some concluding remarks are described.

2 A Superimposed Layer Enhanced by Orientation Filters

The performance of gradient-based pedestrian detector is sensitive to the way in which how large contrast between edges and smooth regions. Therefore, before computing the gradients in raw images, we improve the contrast between edges and non-edge regions by employing the orientation filters. As shown in Fig. 1, Canny edge detector [6] is first applied to the original image and then the orientation filters are used to derives desired edges from the edge map.

The filtering on the edge mask is performed using broadly tuned, steerable, separable filters based on the second derivative of a Gaussian, G_2, and their corresponding Hilbert transforms, H_2 [8], with responses pointwise rectified (squared) and summed. Mathematically, spatiotemporal oriented energies, e, can be computed by

$$e(\boldsymbol{x};\theta)=\left[G_2(\theta)*I(\boldsymbol{x})\right]^2+\left[H_2(\theta)*I(\boldsymbol{x})\right]^2, \tag{1}$$

where $\boldsymbol{x} = (x, y)$, θ is an orientation in a 2D space specified by two polar angles, I is the image sequence, and * denotes the convolution. The initial measure of local

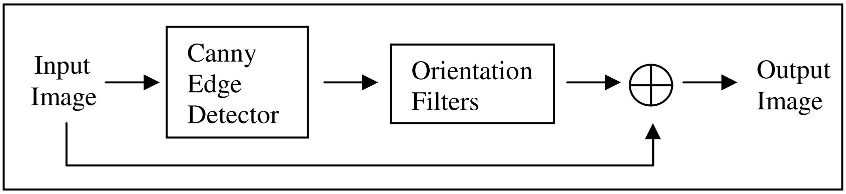


Fig. 1. The workflow of the enhanced layer computation

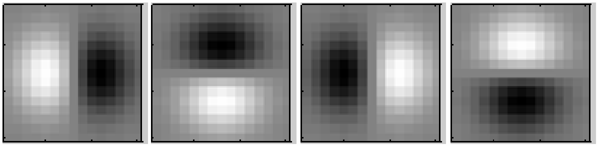


Fig. 2. Steerable Orientated Filters (*0^o, 90^o, 180^o, 270^o*)

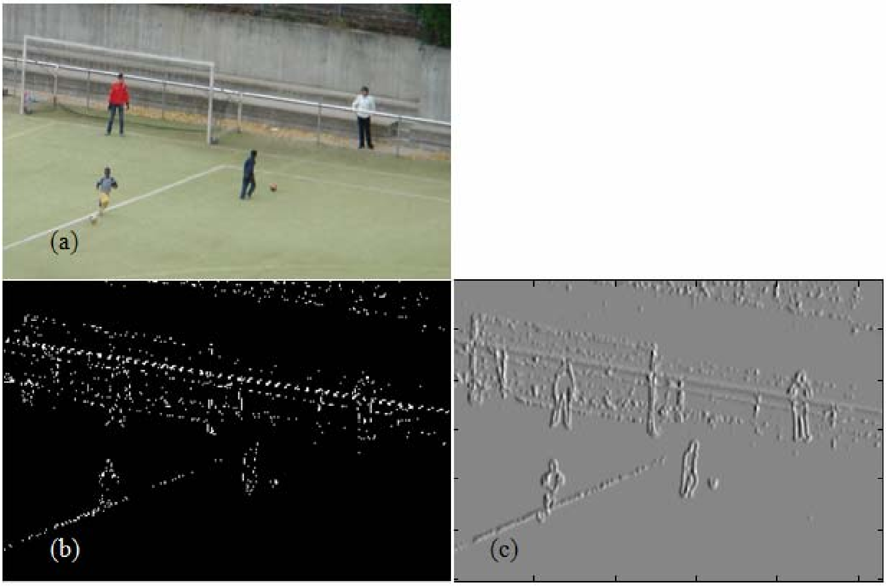


Fig. 3. Demonstration of an image filtered by an orientation filter *0^o*. (a) Original image; (b) Edges detected by Canny edge detector [5]; (c)a filtered image.

energy derived by (1) is dependent on the image's contrast, but an energy measure that is less affected by contrast can be obtained through normalization as follows:

$$\hat{e}(\boldsymbol{x};\theta) = e(\boldsymbol{x};\theta) \Big/ \left(\sum\nolimits_{\tilde{\theta}} e\left(\boldsymbol{x};\tilde{\theta}\right) + \varepsilon\right), \tag{2}$$

where ε is a small bias term that prevents instability when the overall energy content is small. The summation in the denominator covers all orientations at which filtering is performed. In this work, oriented energies are computed at four 2D orientations that correspond to upward, downward, leftward, and rightward motion. An example is shown in Fig. 2. Only four orientations are selected in order to keep the computational requirements to a minimum.

An example is shown in Fig. 3. Fig.3(b) is obtained by using Canny edge detector and Fig.3(c) depicts the edges filtered after applying the orientation filter with angle *0^o*. We can observe that the vertical edges in the edge map are all enhanced, especially the regions containing pedestrians. Therefore, the filtered edge map is combined with the original image for the following feature extractions.

3 AdaBoost Learning

AdaBoost, proposed by Freund and Schapire [2-3], provides a simple and effective approach for stage-wise learning of a nonlinear classification function. Therefore, in this work, AdaBoost is employed as a training approach. A brief introduction of AdaBoost is described as follows. The discrete version of AdaBoost defines a strong binary classifier *H*

$$H(z) = \operatorname{sgn}\left(\sum_{l=1}^{T} \alpha_l h_t(z) \right), \tag{3}$$

using a weighted combination of *T* weak learners h_l with weights α_l. At each new round t, AdaBoost selects a new hypothesis hl that best classifies training samples with high classification error in the previous rounds. Each weak learner *h*

$$h(z) = \begin{cases} 1 & if\ g(f(z)) > 0 \\ -1 & otherwise \end{cases} \tag{4}$$

Explore any feature *f* of the data *z*. In the context of visual object recognition, it is attractive to define f in terms of local features and then use AdaBoost for selecting features by maximizing the classification performance. This kind of approach is first proposed by Viola and Jones [4] who used AdaBoost to train an efficient face detector by selecting a discriminative set of local Haar features.

4 Histogram Features

Local histograms provide effective means to represent visual information for recognition. Similar to [1], to avoid a-priori selection of histogram regions, we consider all rectangular sub-windows *r* of the object. For image regions *r*, we compute weighted histograms of gradient orientations

$$\gamma(x, y) = \arctan \frac{L_x(x, y)}{L_y(x, y)}, \; L_\kappa = I * \frac{\partial}{\partial_\kappa} \left(\frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2} \right), \tag{5}$$

using Gaussian derivatives L_x, L_y defined on the image *I*. γ is discretized into four orientation bins and histograms are incremented by the values of the gradient magnitude $\left\| \left(L_x^2, L_y^2 \right) \right\|_2$. The histograms are normalized to the sum value 1 to reduce the influence of illumination. To preserve some positional information of measurements within the region, regions are divided into sub-regions and we computes histograms separately for each sub-region. Four types of image features, shown in Fig. 4, are then defined for each sub-region *r* by concatenating sub-histograms into feature vectors. All histogram features are computed efficiently using integral histograms [10] which enables real-time consideration of the detection method. During training, we compute features for the normalized training images and apply AdaBoost to select a set of features and hypotheses for optimal performance of classification.

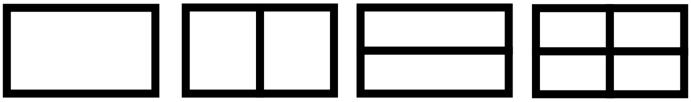


Fig. 4. Sub-region histogram features

5 Experimental Results

To evaluate the performance of our proposed pedestrian detector, the INRIA dataset is adopted for testing. Since our focus is on the investigation of superposing an enhanced layer to the original image, the performance comparison mainly concerns the qualitative evaluation to see the detected results leveraged by the enhanced layer. In Fig. 5, the left side shows the pedestrians detected by using Laptev'06 and the right side demonstrates the ones obtained by applying our proposed approach. To show the detected pedestrians which are of different confidence, the confidence value is

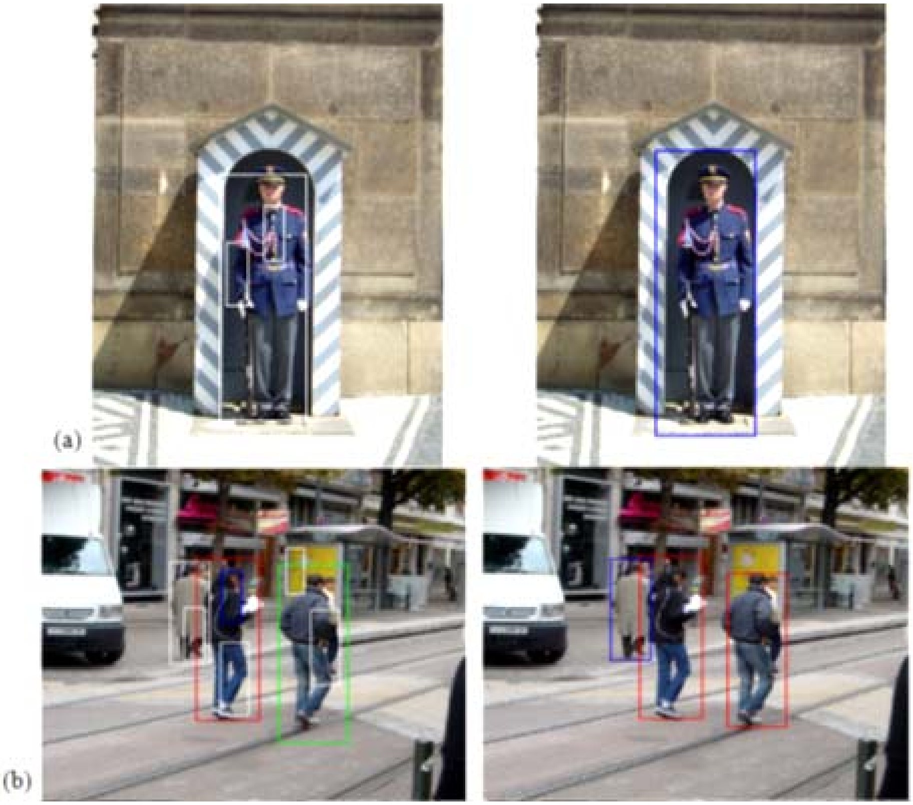


Fig. 5. Demonstration of performance comparison using INRIA dataset. Figures on the right are obtained by our approach and the figures on the left are obtained by Laptev'06 [1].

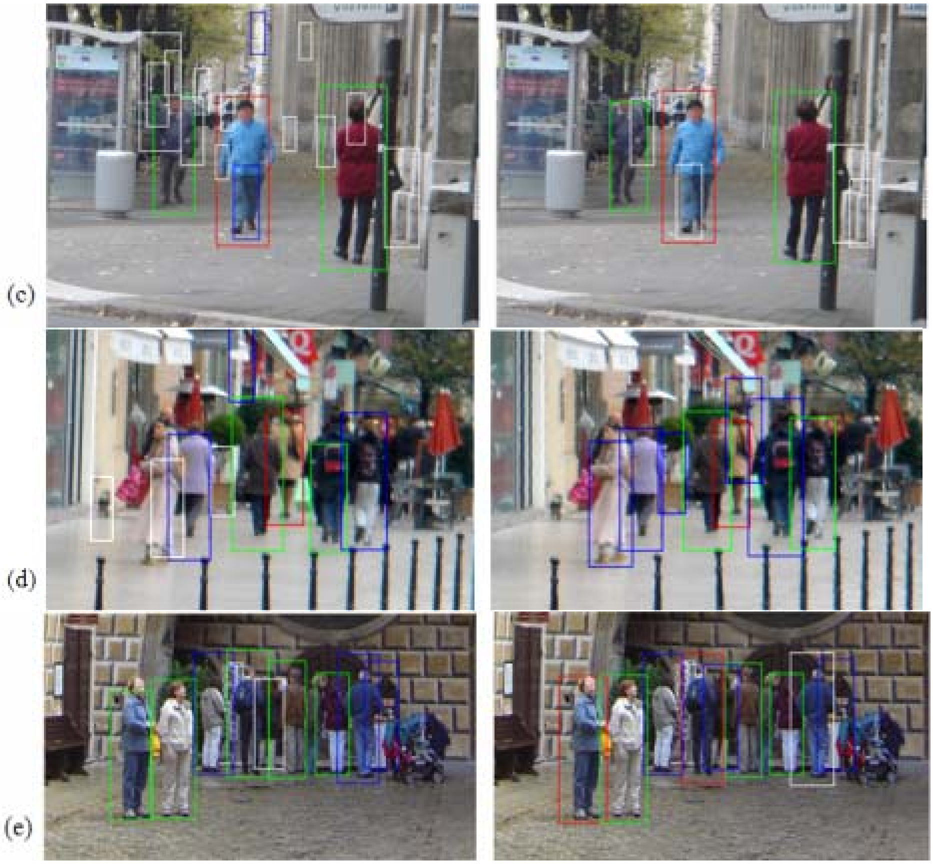


Fig. 5. (*contiued*)

quantized into four levels. In Fig. 5, different colors denote different detecting confidence. In the order of highest to lowest confidence, the colors are red, green, blue and white, separately. The confidence value corresponding to the white bounding boxes is very close to zero. In Figs. 5(a)-5(b), using our method, all the noises are filtered out and the confidence value of the pedestrian is also increased. Besides, the bounding rectangle is more fit to the target than the Laptev'06. In Fig. 5(c), even the confidence value of the real targets is not increased, but most noises are removed. In Fig. 5(d), pedestrians on the right and left are all detected with confidence value increased. Especially, while the left target is almost regarded as noise in Laptev'06, the target is correctly detected with high confidence. In addition, two targets on the right are all precisely enclosed with increased confidence value. In Fig. 5(e), the results are more encouraging. Since the overall scene is more complicated and the pedestrians are occluded by each other, we can still improve the detecting performance.

For more performance comparison, while the lack of the source code, we demonstrate the qualitative evaluation by an example with comparing to Laptev'06 and Chen'08 [6] in Fig. 6. In Chen's method, we can observe that each enclosing bounding rectangle is loosely coupled with the detected target. Our proposed approach can

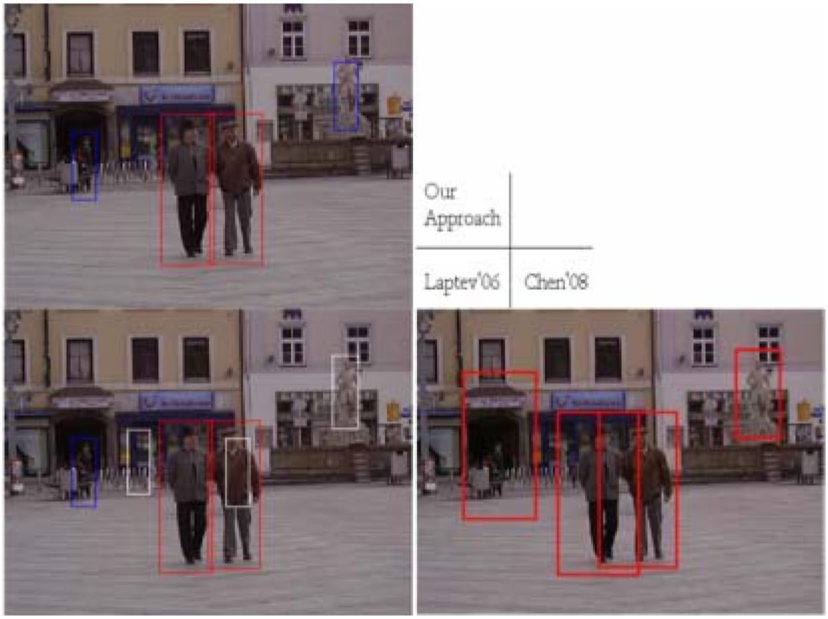


Fig. 6. Demonstration of performance comparison. (a) Our proposed approach; (b) Laptev'06 [1]; (c) Chen'08 [6].

derive more tightly coupled bounding rectangle with the targets. Comparing to the Laptev's method, our approach can eliminate most noises and at the same time increase the confidence value of the targets. Therefore, our approach demonstrates the robust detecting results and outperforms over the methods of Laptev'06 and Chen'08.

6 Conclusion

In this work, we have proposed a method that can detect pedestrians in a single image based on the combination of AdaBoost learning with a local histogram features. Besides, instead of using the raw image for further processing, we have introduced a layer enhanced by orientation filters which are superimposed to the original image. As the edge features are properly improved by the layer, the gradient-based descriptors have more discriminative power over the ones without any enhanced layers. Experimental results have shown the superior performance of our method and thus demonstrate its robustness with the novel enhanced layer embedded pedestrian detector.

References

1. Laptev, I.: Improvements of Object Detection Using Boosted Histograms. In: Proc. of British Machine Vision Conference (2006)
2. Freund, Y., Schapire, R.E.: A Decision-Theoretic Generalization of On-Line Learning and An Application to Boosting. Journal of Computer and System Sciences 55(1), 119–139 (1997)

3. Schapire, R.E., Singer, Y.: Improved Boosting Algorithms Using Confidence-Rated Predictions. Maching Learning 37(3), 297–336 (1999)
4. Viola, P., Jones, M.: Rapid Object Detection Using A Boosted Cascade of Simple Features. In: Proc. of IEEE Computer Vision and Pattern Recognition, pp. 511–518 (2001)
5. Canny, J.: A Computational Approach to Edge Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 8(6) (November 1986)
6. Chen, Y.T., Chen, C.S.: Fast Human Detection Using A Novel Boosted Cascading Structure with Meta Stages. IEEE Transactions on Image Processing 17(8) (August 2008)
7. Dalal, N., Triggs, B.: Histograms of Oriented Gradients for Human Detection. In: Proc. of IEEE Computer Vision and Pattern Recognition (2005)
8. Freeman, W., Adelson, E.: The Design and Use of Steerable Filters. IEEE Pattern Analysis and Machine Intelligence 13(9), 891–906 (1991)
9. Viola, P., Jones, M., Snow, D.: Detecting Pedestrians Using Patterns of Motion and Appearance. In: Proc. of IEEE Computer Vision and Pattern Recognition, vol. 2, pp. 734–741
10. Porikli, F.M.: Integral Histogram: A Fast Way to Extract Histograms in Cartesian Spaces. In: Proc. of IEEE Computer Vision and Pattern Recognition, vol. 1, pp. 829–836 (2005)
11. Chen, D.Y., Cannons, K., Tyan, H.R., Shih, S.W., Mark Liao, H.Y.: Spatiotemporal Motion Analysis for the Detection and Classification of Moving Targets. IEEE Trans. on Multimedia 10(8), 1578–1591 (2008)

Mining Influential Bloggers: From General to Domain Specific

Yichuan Caiv and Yi Chen

Arizona State University P.O. Box 878809, Tempe, AZ 85287-8809, USA
{yichuan.cai,yi}@asu.edu

Abstract. With rapid development of web 2.0 technology and e-business, bloggers play significant roles in the whole blogosphere as well as the external world. Specially, the most influential bloggers can bring great business values to modern enterprise in multiple ways, by increasing market profits and enlarging business impacts. The bloggers' influences can be deployed only in a specific domain, e.g. computer companies only can utilize the influence bloggers' expertise in computer knowledge, not their expertise in modern art or others. Despite that several influential bloggers mining systems are available, none of them consider the domain specific feature and their evaluations are based on generic influence, which is not applicable for real application requirements, such as business advertisement, personalized recommendation and so on. In this paper, we propose an effective model to mine the top-k influential bloggers according to their interest domains and network proximity. We investigate an effective algorithm to evaluate a blogger's influence and develop a domain specific influential blogger mining system. The experiment results show that our system can effectively mine influential bloggers and is applicable to diverse applications.

1 Introduction

With the advantage of the modern technology, Web 2.0 provides a second generation web-based communities with services such as forums, wikis, blogs, folksonomies and etc., which can facilitate the communication, collaboration, and information sharing among web users. Blogs, as one of the most important components of web 2.0 services, provide a conductive platform for web bloggers to post their logs of events and share their personal insights with the blog visitors, and let them to read and write down feedbacks. By building such a virtual communities, blogs have attracted great interests from web users, industry, as well as research communities and become one of the most popular and widely used web 2.0 services.

In light of that blog readers are likely to be influenced by the bloggers, an increasing number of corporations now start to use blogs as a new product marketing strategy to enlarge their profits by leveraging influential bloggers which have a large population of potential readers. The main reasons come from the following perspectives.

First, the posts from an influential blogger often have a larger impact on their readers' purchasing decision than advertisements from the company, since people typically trust and act on recommendations from knowledgeable people and their friends. To carry out promotional effort, using a "word-of-mouth" advertising - marketing the bloggers with

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 447–454, 2009.

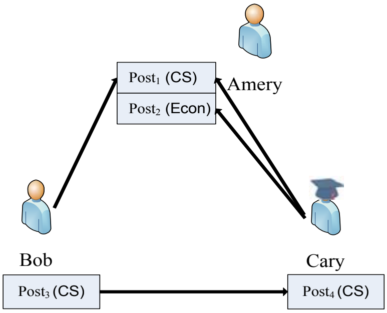


Fig. 1. A Sample "Influence Graph" in Blogsphere

strong influence and leveraging the customers themselves to be unofficial spokesmen, a product can be marketed in a much more cost effective way than traditional methods. Some researchers have given the theoretical explanation of the how the influential members have the influence on the other members' action in the community. [4] investigates causes of user action correlation, which could be categorized into three types: the influence, the homophily and environment. Second, communicating with influential customers and analyzing their blogs can bring companies good understanding and insights of the key concerns and new trends of customers' interest for product improvements with much less cost compared with searching, aggregating and analyzing all the relevant blogs. The influential ones' solutions and suggestions are often very valuable due to the sense of authority they possess.

Due to these potential business opportunities, recently, identifying top-k influential bloggers begins to attract more and more research interests [10,3,5]. They measure the influence among bloggers based on the "post-reply" relationships, which are modeled in an *influence graph*.

Let's take a look at a sample *influence graph* in Figure 1, Amery has two posts, which are $post_1$ with comments from Bob and Cary, and $post_2$ with comments from Cary. Assume that $post_1$ discusses some programming skills in computer science, while $post_2$ investigates the recent economic depression and possible trends in the next couple of months. To evaluate the influence of Amery, existing work [3] considers factors such as the number of inlinks/outlinks, the number of comments as well as the length of those comments, to measure the influence degree of blogger Amery.

Although it is a straightforward metric to use, some important and valuable information embedded in the "post-reply" relationship are ignored by the existing works. First, the two posts of Amery are related to different domains, when evaluating the influence of Amery on computer science or economics, it is necessary to consider them separately. In another word, the influence of a blogger is domain specific, consequently a good model should capture this information. Secondly, the influence of each comment may have different impact power, depending on who issues it. For example, Cary is an expert in computer science, while Bob is an entry level freshman of computer science. Their comments on Amery's $post_1$ should be treated differently, and it is easy to see Cary's comment would enhance the influence of $post_1$ more. Thirdly, the comments from other bloggers could be positive, negative or neutral, and these sentimental factors also affect the post's influence among commenters.

Furthermore, to evaluate the influence of a blogger, only considering "post-reply" relationship is not enough since people may not put comments on others' blogs even s/he has great interest in it. Also, an influence blogger would usually have many external links to his/her blog, e.g.when people find someone's blog is very interesting, s/he may directly add a link to his/her own space. External links to a blogger provides another possible way to measure the influence of that blogger. PageRank [8] and HITS [6] are effective algorithms to evaluate the authority of a page (and link). We also take the authority factors into consideration when we evaluate the influence of the bloggers.

In this paper, we propose a mechanism that can investigate the top-k domain specific influential bloggers. Considering both "post-reply" relationship and general link information of a blogger, it is possible to evaluate the comprehensive influence of a blogger.

There are several challenges in the blogger's domain specific influence evaluation. For instance, what should be considered when evaluating a blogger's influence in a specific domain? Should his topology proximity be included in the evaluation? How to capture the quality of a post, and measure its quality? How to judge the degree of the impact of a blogger on a commenter? How to evaluate a blogger's influence on different domains? In the following sections, we investigate these issues in turn and explore our model and system in detail.

The rest of paper is organized as follows: we introduce our interest vector model and discuss how to evaluate the domain specific influence in Section 2. Experimental studies are presented in Section 3 and finally Section 4 concludes the paper.

2 Our Approach

In this section, we will first introduce an interest vector model which represents the domain specific interest for a blogger, and then we discuss how to evaluate the blogger's domain specific influence in detail. At last, we will illustrate how to apply our model in different applications accordingly.

2.1 Interest Vector Model

A blogger's blog space is composed of multiple posts, and each of them can belong to one or multiple possible domains, e.g. some posts of latest NBA news are belong to the `sports` domain, the broadcast of U.S. present Obama's speech on ASU's commencement belongs to the `politics` domain as well as the `education` domain. In order to describe a blogger's interest on a specific domain, we define an "interest vector model", which represents the possibility of a post belonging to a specific domain, as follows:

Given a blogger b_i's post d_k, its interest can be quantified as a vector in the whole interest space, called *interest vector*:

$$IV(b_i, d_k) = \{iv_1, iv_2, ..., iv_N\}$$

where $iv_t \in [0, 1], (1 \leq t \leq N)$. t is the t^{th} dimension in the interest space, and N is the total number of domains, which can be predefined according to some standard categories(such as Open Directory Project) [1].

[1] http://www.dmoz.org/

Most blog service providers can support predefined categories to users, e.g. MSN Space allows bloggers to select a category from a candidate list for his/her new post, we denote them as "category tag" for the post. However, the coverage of "category tag" is pretty low: according to our sample of 1000 bloggers in MSN spaces in the empirical study, we find about 70% of the posts do not explicitly provide their "category tag".

When a tag information is not available, we mine the blogger's interest information from their post content and construct the *interest vector*. Then we use the naïve Bayesian classifier [2] to get the possibility that a post belongs to certain predefined category. Formally, a post's interest vector is calculated as follows:

$$iv(b_i, d_k, C_t) = \frac{P(b_i, d_k|C_t)P(C_t)}{\sum_{n=1}^{N}(P(b_i, d_k|C_n)P(C_n))} \tag{1}$$

where $P(C_t|b_i, d_k)$ is the possibility that blogger b_i's post d_k belong to category C_t. $iv(b_i, d_k, C_t)$ is the possibility that the blogger b_i's post d_k belong to category C_t normalized by the summation of the possibility that blogger b_i's post d_k belongs to all the predefined categories.

2.2 Influence Evaluation

A blogger's influence on a specific domain can be treated as one component of his/her overall influence, hence before we investigate the domain specific influence for each blogger, it is necessary to quantify each blogger's personal overall influence. Intuitively, an influential blogger always has high quality posts together with many commenters, and high authority in the whole network. The posts and comments reflect a blogger's expertise and popularity, while the authority reflects his/her position in the whole network and linkage with other bloggers. Correspondingly, the overall influence of a blogger can be measured by two parts: the summation of his/her posts' influence noted as `Accumulated Post (AP)` influence score, and his/her authority in the network, noted as `General Links (GL)` influence score. Since each of his/her post reveals his/her interest in specific domains, we choose "post" as our basic analysis unit, rather than a blogger. The GL is similar to a webpage authority and PageRank [8] value becomes a natural choice of the approximation of the authority. The GL score of each blogger can be calculated by standard Page Rank algorithm or direct approximated by certain Page Rank Value provider.[2] Hence, we can define the personal overall influence of a blogger as following:

$$Inf(b_i) = \alpha * \sum_{k=1}^{|P_i|} Inf(b_i, d_k) + (1-\alpha) * GL(b_i) \tag{2}$$

where $\sum_{k=1}^{|P_i|} Inf(b_i, d_k)$ is AP score, and $|P_i|$ is b_i's total number of posts. Specifically, $Inf(b_i, d_k)$ is influence score of blogger b_i's post d_k (which will be introduced later). $GL(b_i)$ is the GL score, α is the parameter to tune the relative importance of AP score and GL score.

[2] "Cubestat (http://www.cubestat.com)"

The post's influence score is always reflected by its quality and the comments on the post. In order to define the score of each individual post d_k of blogger b_i, $Inf(b_i, d_k)$ in Eq. 2, we consider both the quality of the post's content and the commenters' impact.

$$Inf(b_i, d_k) = \beta * QualityScore(b_i, d_k) + \quad (1 - \beta) * CommentScore(b_i, d_k) \tag{3}$$

where β is a parameter used as a weight for the two parts.

$QualityScore(b_i, d_k)$, as the first component of $Inf(b_i, d_k)$, is evaluated by the length of a post in existing works [3]. The longer, the more influence the post has. Besides, the novelty plays an important role. The more novel the post is, the more influence it has. The novelty of a post reflects the creativity of that post, whether b_i's post d_k is his original idea or carbon copy from others. In our setting, the $QualityScore(b_i, d_k)$ is the product of a post's length and novelty. The novelty is a numeric value between 0 and 1, which could be mined from the post content.

$CommentScore(b_i, d_k)$, the second component of a post's influence, reflects the impact on the commenters. Each comment's score, is proportional to the summation of the commenter blogger's personal overall influence score $Inf(b_j)$ and his/her attitude toward the post, which is the sentiment factor $SF(b_i, d_k, b_j)$ of b_j's comment to b_i's post d_k. Also, one commenter may put multiple comments on other blogger's spaces, and his/her impact to peers will be shared, hence we normalize the score by the total comments $TC(b_j)$ of commenter b_j. The $CommentScore$ is defined as following:

$$CommentScore(b_i, d_k) = \sum_{j=1}^{|b_i, d_k|} \frac{Inf(b_j) * SF(b_i, d_k, b_j)}{TC(b_j)} \tag{4}$$

$|b_i, d_k|$ is the total number of comments on blogger b_i's post d_k. The sentiment factor $SF(b_i, d_k, b_j)$ captures the commenter's attitude, and can be classified into three categories: positive, negative or neutral. We use the following heuristics to predict its value: as long as a comment contains certain positive words (such as “agree”, “support”, “conform”, which are positive comments), we treat it as a positive comment. If it contains negative words (such as “disagree”, “hate”, “not credible”), we treat it as a positive comment. Otherwise, we treat it as a neutral comment.

From Eq.3 and Eq. 4, we get the following equation:

$$Inf(b_i, d_k) = \beta * QualityScore(b_i, d_k) + (1 - \beta) * \sum_{j=0}^{|b_i, d_k|-1} \frac{Inf(b_j) * SF(b_i, d_k, b_j)}{TC(b_j)} \tag{5}$$

Each blogger's post has an equation in the format of Eq. 5 , and all the equations of bloggers' post can be solved by iterative method efficiently [1]. The solution to the equation set provides the influence score for each blogger's post, from which we can get each blogger's total influence score by using the Eq. 2.

2.3 Domain Specific Influence Score

Finally, we evaluate the blogger's influence score for each domain. There are several predefined domains which are very popular in the blogosphere, such as Travel, Art, Sports, etc. Intuitively, the post's domain influence score w.r.t certain domain is proportional to the post's total influence score and the $iv(b_i, d_k, C_t)$ (interest score of b_i's post d_k's belongs to predefined category C_t), which is evaluated as following:

$$Inf(b_i, C_t) = \sum_{k=1}^{|P_i|} Inf(b_i, d_k) * iv(b_i, d_k, C_t) \tag{6}$$

$iv(b_i, d_k, C_t)$ could be calculated by existing interests mining method [7,9], and we choose naïve Bayesian algorithm [2] in our implementation.

Example 1. Let us consider the influence graph in Figure 1 as our example and explore the whole domain specific influence score calculation in detail. To be concise, we use the first character of blogger's name to represent the blogger, e.g A stands for $Amery$. We can get the following equations for the set of bloggers according to Eq. 5:

$$Inf(A, P_1) = \alpha * QualityScore(A, P_1) + \frac{1}{2} * \beta * ((\alpha * Inf(B, P_3) + (1-\alpha) * GL(B)) * SF(A, P_1, B))) + \frac{1}{2} * \beta * ((\alpha * Inf(C, P_4) + (1-\alpha) * GL(B)) * SF(A, P_1, C)))$$

$$Inf(A, P_2) = \alpha * QualityScore(A, P_2) + \frac{1}{2} * \beta * ((\alpha * Inf(C, P_4) + (1-\alpha) * GL(C)) * SF(A, P_2, C)))$$

$$Inf(B, P_3) = \alpha * QualityScore(B, P_3)$$

$$Inf(C, P_4) = \alpha * QualityScore(C, P_4) + \frac{1}{2} * \beta * ((\alpha * Inf(B, P_3) + (1-\alpha) * GL(B)) * SF(C, P_4, B)))$$

As we can see from these equations, $Inf(A, P_1)$, $Inf(A, P_2)$, $Inf(B, P_3)$ and $Inf(C, P_4)$ are variables, after we solve these equations, it is easy to get the values of $Inf(A)$, $Inf(B)$ and $Inf(C)$ which are the bloggers overall influence scores. Based on them, we can further approach to domain influence scores $Inf(A, Econ)$, $Inf(B, CS)$ and $Inf(C, CS)$ for domain {Econ, CS} directly.

3 Experiments

In order to evaluate the effectiveness of our method, we conduct comprehensive analysis on real data set in the following subsections. We use Microsoft MSN space [3] as our test data set, which is one of the most popular blog service providers. Each blogger can write posts on their own blogs and leave comments on others' posts. We have crawled around 1000 MSN spaces with user profiles, comments and their recent posts.

We predefine ten interest domains as following: {Traveling, Computer, Communication, Education, Economics, Military, Sports, Medicine, Art, Politics}, because these domains cover most interests of bloggers. When calculate the Genaral Link(GL) influence score, we found that most bloggers' PageRank values are very small. In our data set, 90% blog's PageRank value is less than 1, about 99% blogs' PageRank value is less than 3. Instead of using PageRank directly, we utilize Microsoft Live Indexed Pages as the approximation of PageRank value, which could be obtained from the website "Cubestat"(http://www.cubestat.com/).

3.1 Domain Specific or Not

To evaluate the effectiveness of our model, we invite 10 users to do a user study, who compare the recommendation performance of top 3 influential bloggers mined from

[3] http://home.spaces.live.com/

Table 1. User Evaluation of Average Applicable Scores for Influential Bloggers (General VS. Domain Specific)

Average Applicable Scores	Traveling	Art	Sports
General	3.2	3.4	3.2
Domain Specific	4.4	4.0	4.6

general domain and specific domains (Traveling, Art, and Sports). For the top 3 bloggers in the general and domain-specific lists, we send the URL of each blogger to end users, and ask users to score them from 1 to 5 according to their understanding of a specific application scenario, e.g. “Suppose you are the sales manager in Nike, which blogger will you choose to send your advertisement to? ”. The results of average applicable score for the user study is shown in Table 1. As we can see, our model has better evaluation results than that of general influential blogger recommendation system cross different domains. Especially, the Sports domain has a much higher evaluation score 4.6 than that of general one of 3.2.

3.2 Impact of Weighting Parameters

As we have discussed in Section 2, α is the parameter to tune the related importance of accumulated post influence and general link influence and β is the parameter to adjust the related importance of each post's quality score and comment score. To see the impact of these tuning parameters, we randomly choose the Art domain as an example.We tune parameters α and β by fixing one and changing the other, observe the variance of the ranking results. For space limit reason, we only show sample results in Table 2 with β changing by fixing α=0.5.

Table 2. Top 5 influential bloggers with different β given α

$\beta = 0.9$	$\beta = 0.5$	$\beta = 0.1$	$\beta = 0.01$	$\beta = 0$
youyou	youyou	youyou	youyou	kelly
sky	sky	sky	kelly	winson
newwishes	newwishes	sabrina	sky	best
sabrina	sabrina	newwishes	sabrina	whenlove
Frank	Frank	kelly	newwishes	youyou

As we can see from Table 2, the top 1 influential blogger is changed from “youyou” to “ kelly”. Take a close look at these two bloggers' posts, “youyou” has many high quality posts, together with a lot of positive comments from high influential commenters. While for “kelly”, whose posts always reproduce from other sources without rich contents, but have many influential commenters, possibly from her friends on the blogosphere. No matter the quality of her blog's posts are good or not, she always has a high comment score. This example shows the relationship between the two components of a post, and both of them have impacts on the evaluation of a post's influence score.

4 Conclusions and Future Work

In this paper, we address a novel problem of identifying influential bloggers considering domain specific information. To better identify influential bloggers, we analyze the interests of bloggers, evaluate the influence of a blogger according to their interest domains. The evaluation with data from a real world blog site, Microsoft MSN space, shows the effectiveness of our approach.

In the future, we will further extend our system to visualize the influential bloggers and cooperate with real business applications.

Acknowledgement

The authors are partially supported by NSF CAREER award IIS-0845647 and NSF grant IIS-0740129. We would like to thank Qihong Shao for her valuable input to the paper.

References

1. http://en.wikipedia.org/wiki/gauss-seidelmethod
2. http://en.wikipedia.org/wiki/naivebayesclassifier
3. Agarwal, N., Liu, H., Tang, L., Yu, P.S.: Identifying the influential bloggers in a community. In: WSDM (2008)
4. Anagnostopoulos, A., Kumar, R., Mahdian, M.: Influence and correlation in social networks. In: SIGKDD, pp. 509–516. ACM Press, New York (2008)
5. Kempe, D., Kleinberg, J., Tardos, É.: Maximizing the spread of influence through a social network. In: KDD 2003. ACM Press, New York (2003)
6. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. Journal of the ACM
7. Liu, Y., Liu, W., Jiang, C.: User interest detection on web pages for building personalized information agent. Innovations in Information Technology (2006)
8. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project (1998)
9. Paik, W., Yilmazel, S., Brown, E., Poulin, M., Dubon, S., Amice, C.: Applying natural language processing (nlp) based metadata extraction to automatically acquire user preferences. In: Proceedings of the 1st international conference on Knowledge capture (2001)
10. Scrpps, J., Tan, P.-N., Abdol-Hossein: Node roles and community structure in networks. In: WEBKDD (2007)

Efficiency of Node Position Calculation in Social Networks

Piotr Brodka, Katarzyna Musial, and Przemyslaw Kazienko

Wroclaw University of Technology, Institute of Computer Science,
Wybrzeze Wyspianskiego 27, Wroclaw, Poland
{piotr.brodka,katarzyna.musial,kazienko}@pwr.wroc.pl

Abstract. Social network analysis offers many measures, which are successfully utilized to describe the social network profile. One of them is node position, useful to assess the importance of a given node within both the whole network and its smaller subgroups. However, to analyze large social networks a lot of effort and resources are necessary. In this paper, some algorithms that can be utilized in the process of node position evaluation are presented and their efficiency is tested. In particular, three distinct algorithms were developed and compared: PIN Edges, PIN Nodes, and PIN hybrid.

Keywords: node position, social network analysis, PIN algorithm, calculation efficiency.

1 Introduction

Social networks attract more and more researchers attention. With the growth of social networks popularity, the greater and greater number of methods have been developed to investigate and analyze this kind of networks. One of the crucial issues in social network analysis is the problem of extracting of the most important (central) members. There are several methods used for this purpose, such as node position [8], [9], [10], rank prestige [15], indegree centrality [15], [1], outdegree centrality [12], [14], closeness centrality [2], [13], proximity prestige [15], betweenness centrality [5], [6], [7], and others. For each of them the appropriate algorithms were proposed [3]. Although the variety of algorithms exists, there is lack of the research on their efficiency. The efficiency tests and complexity analysis [4] enable to select proper algorithms fast enough to compute centralities within the large network.

2 Node Position Measure

Node position *NP* is the centrality measure studied in this paper. It enables to estimate how valuable the particular individual within the human community is [8], [9], [10]. In other words, the importance of every member can be assessed by calculating their node position. In general, the greater node position one

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 455–463, 2009.

possesses the more valuable this member is for the entire community. It is often the case that we only need to extract the highly important persons, i.e. with the greatest node position. Such people are likely to have the biggest influence on others. As a result, we can focus our activities like advertising or target marketing solely on them and expect them to entail their acquaintances.

Let us consider the weighted social network $SN(M, R)$, where M is the set of network members and R — the set of their relationships. The importance of member $x \in M$ in $SN(M, R)$, is expressed by the node position function, tightly depends on the strength of the relationships that this individual maintains as well as on the node positions of their acquaintances, i.e. the first level neighbors. In other words, the member's node position is inherited from others but the level of inheritance depends on the activity of the members directed to the considered person, i.e. the intensity of common interaction, cooperation or communication. The activity contribution of one user absorbed by another is called commitment function. Node position $NP(x)$ of individual x respects the values of node positions of the direct x's acquaintances as well as their activities in relation to x. Node position is calculated in the iterative way, i.e. $NP_{n+1}(x)$ results from the previous node positions $NP_n(y_i)$ of neighbors y_i, as follows:

$$NP_{n+1}(x) = (1 - \varepsilon) + \varepsilon \cdot \sum_{i=1}^{m_x} (NP_n(y_i) \cdot C(y_i \rightarrow x)) \tag{1}$$

where: y_i — x's acquaintances, i.e. the members who are in direct relationship to x; m_x — the number of x's nearest acquaintances. ε – the constant coefficient from the range $[0; 1]$, which denotes the openness to the external influence; $C(y_i \rightarrow x)$ – the function that denotes the contribution in activity of y_i directed to x; $NP_{n+1}(x)$ and $NP_n(x)$ — the node position of member x after the $n + 1$st and nth iteration, respectively.

To perform the first iteration, any initial values of node position $NP_0(x)$ need to be assigned to all $x \in M$. Since the calculations are iterative, we also need to introduce a stop condition. For this purpose, a fixed precision coefficient τ is used. Thus, the calculation is stopped until the following criterion is met: $(x \in M)$ $|NP_n(x) - NP_{n-1}(x)| \leq \tau$. Obviously, another version of the stop condition can be also applied: $|SNP_n - SNP_{n-1}| \leq \tau$, where: SNP_n and SNP_{n-1} — the sum of all node positions after the nth and $n - 1$th iteration, respectively.

3 Position in Network Algorithms

Based on Eq. 1, the PIN algorithms (**P**osition **I**n the **N**etwork) in three different versions were developed, i.e. PIN^{nodes}, PIN^{hybrid}, and PIN^{edges}. These algorithms differ in the implementation and in consequence their efficiency varies. All algorithms require the same set of input data and provide as the output the social position values for each network member together with the number of iterations required to meet the given stop condition.

```
Input:
  M, R - set of members and their relationships,
  C - list of commitment values, one for each ordered pair $(x_1, x_2) \in M$,
  $NP_0 =< NP_0(x_1), NP_0(x_2), \cdots, NP_0(x_m) >$ - vector of initial node positions,
  $\varepsilon \in [0;1]$ - coefficient from Eq. 1,
  $\tau$ - stop condition (precision coefficient), e.g. $\tau := 0.00001$.
Output:
  $NP =< NP(x_1), NP(x_2), \cdots, NP(x_m) >$ - vector of final node positions,
  n - the number of iterations,

begin
  $n := 0$;
  $NP_{prev} := NP_0$; $NP := NP_0$;
  divide $M$ into $m$ disjunctive subsets $\{s_1, \cdots, s_m\}$ /* used in PIN_Hybrid */
  repeat
    PIN_Nodes(); /* invoke here proc.: PIN_Edges() or PIN_Hybrid() */
    $n := n + 1$;
  until stop condition $\tau$ is fulfilled for all members;
end

procedure PIN_Nodes() begin
  for (each member x from M) do begin
    $NP[x] := (1 - \varepsilon)$;
    for (each member y from M) do
      $NP[x] := NP[x] + \varepsilon \cdot NP_{prev}[y] \cdot C[y, x]$;
  end
  $NP_{prev} := NP$;
end

procedure PIN_Edges() begin
  for (each edge r(x, y) from R) do
    $NP[y] := NP[y] + NP[x] \cdot C[x, y]$;
  for (each member x from M) do
    $NP[x] := (1 - \varepsilon) + \varepsilon \cdot NP[x]$;
end

procedure PIN_Hybrid() begin
  for(each disjunctive subset $s_k$) do
    for(each edge $r(x, y)$ where $y$ is member of $s_k$) do
      $NP[y] := NP[y] + NP[x] \cdot C[x, y]$;
  for(each member $x$ from $M$) do
    $NP[x] := (1 - \varepsilon) + \varepsilon \cdot NP[x]$;
end
```

The first proposed algorithm PIN^{nodes} is the direct, raw implementation of the node position concept, Eq. 1. It has been completed without any optimization techniques. All the calculations are made from so called "node perspective", i.e. the node position is calculated one by one for each network node — member, see procedure *PIN_Nodes*(). First, two lists NP_{prev} and NP that contain the node position values are created. NP_{prev} stores node positions from the previous iteration whereas in NP the final values calculated in the current iteration are preserved. At the beginning, the initial node position values NP_0 are assigned to NP_{prev} and NP - necessary for alternate algorithms. Afterwards, for each member $x \in M$ its SP is set to $(1 - \varepsilon)$. Next, for each member $y \in M$ the value of commitment function $C(y \to x)$ is multiplied by $NP_{prev}[y]$ and ε. The result is added to the current value of x's node position, i.e it is stored in $NP[x]$. Finally, the values from NP are assigned to NP_{prev} and the iteration finishes. The next iteration is performed unless the stop condition is met.

The second developed algorithm is called PIN^{edges} and its all calculations are made from so called "edge perspective", i.e. the node position is calculated rather by taking into the consideration the edges (set R) and their weights (commitment function assigned to the edges), followed by evaluation of node position one by one for each network node — member, see procedure *PIN_Edges*(). For each edge $r(x, y)$ from set R of all edges increase the node position value of user y $(NP(y))$ with the node position of x $(NP(x))$ multiplied by commitment function from user x to y $(C(x \to y))$. Next, for each M's member multiply the obtained node position of the given user by ε and add the appropriate component $1 - \varepsilon$.

The third algorithm, named PIN^{hybrid}, combines both previous approaches, see procedure *PIN_Hybrid*(). All nodes of the network are divided into m disjunctive subsets $\{s_1, s_2, \cdots, s_m\}$. For each subset s_k created, the following action is performed: for each edge $r(x, y)$, which y belongs to subset s_k, increase y's node position $NP(y)$ with x's node position $NP(x)$ multiplied by the value of commitment function from user x to y $(C(x \to y))$. Next, for each member of M multiply the obtained node position by ε and add the component $1 - \varepsilon$.

4 Efficiency Tests

Results of all three methods ware compared separately to each other. Kendalls coefficient of concordance was used [11] for this purpose. Its value was always higher than 0.87, this means that rankings were very similar. Additionally, the order of top 50 users was analyzed - it was the same for each of three methods. The main aim of the performed efficiency tests was to investigate, which of the three developed algorithms: PIN^{nodes}, PIN^{edges} or PIN^{hybrid} is the most efficient. The efficiency tests were split into two main stages. First, the influence of ε coefficient on processing time of different variants of PIN algorithms is investigated. In the second phase, the tests were performed on the networks of different size, i.e. with different number of nodes and edges. These were random networks generated for the purpose of the experiments.

The first part of experiments was performed on the real data received from one telecommunication company. The network consisted of over 4 million users and over 17 million connections. The tests were carried out for several values of ε and for all three algorithms (PIN^{nodes}, PIN^{edges} and PIN^{hybrid}), Tab. 1.

It can be easily noticed that the processing time is the biggest for PIN^{nodes} and the shortest for the PIN^{edges}. The PIN^{edges} algorithm is over 120 times faster than PIN^{nodes} and about 2.5 times faster than PIN^{hybrid}.

Table 1. Average processing time for one iteration in relation to ε coefficient, [s]

ε	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
PIN^{nodes}	338,881	338,589	338,699	338,740	338,761	338,859	338,689	338,409	339,236
PIN^{edges}	2,646	2,740	2,790	2,769	2,800	2,749	2,757	2,722	2,703
PIN^{hybrid}	6,764	6,663	6,651	6,739	6,779	6,775	6,664	6,746	6,679

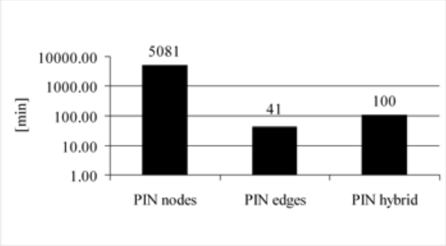


Fig. 1. Average processing time of one iteration for different variants of PIN algorithm

When the ε coefficient is taken into consideration then the average processing time of one iteration for PIN^{nodes} is over 5000 minutes, for PIN^{edges} is around 41 minutes and for PIN^{hybrid} equals 100 minutes, Fig. 1. The analysis of standard deviation of these values enables to assess how the ε coefficient influences the processing time of PIN algorithms. The smallest standard deviation is for PIN^{edges} algorithm and equals 0.79 min whereas the biggest one is for PIN^{nodes} (3.79 min) and this is intuitive because the average time of one iteration is also the biggest. These standard deviations are small in comparison to average processing time of one iteration for different ε values so it can be assumed that the value of ε coefficient does not influence the processing time to a significant extent.

The next stage of the efficiency tests was performed on random social networks and for the fixed value of ε, i.e. $\varepsilon = 0.8$. For each test 25 different random directed networks were generated.

First, the tests were performed for the PIN^{nodes} algorithm, Tab. 2. The processing time for the largest network (100,000 nodes and 100,000 edges) was approximately 1950 times longer than for the smallest one (1,000 nodes, 1,000 edges). It reveals that the network size has the great influence on processing time. The bigger network, the longer processing time.

The similar tests were carried out for the PIN^{edges} algorithm, Tab. 3. The processing time for the largest network (100,000 nodes and edges) was approximately 84 times longer than for the smallest one (1,000 nodes and edges). Hence, the influence of the network size on processing time is much smaller than in case of the PIN^{nodes} algorithm.

The last tests were performed for PIN^{hybrid} algorithm, Tab. 4. The processing time for the largest network compared to the smallest one was approximately 88 times longer. Similarly to PIN^{edges}, it points out that the influence of the

Table 2. Processing time of the PIN^{nodes} algorithm for different network sizes [s]

Nodes Edges	1,000	5,000	10,000	50,000	100,000
1,000	15.54	27.78	39.96	138.77	255.04
5,000	265.50	326.59	392.55	861.69	1444.97
10,000	976.55	1,144.60	1,189.02	2,160.86	3,367.28
50,000	10,538.83	10,937.89	11,241.66	14,701.08	17,517.50
100,000	22,141.31	22,185.78	23,360.89	26917.69	30,304.94

Table 3. Processing time of the PIN^{edges} algorithm for different network sizes [s]

Nodes Edges	1,000	5,000	10,000	50,000	100,000
1,000	0.45	0.73	1.05	3.55	6.12
5,000	1.72	2.10	2.42	5.04	8.13
10,000	3.46	3.81	3.96	6.55	10.09
50,000	16.57	16.20	16.39	18.95	22.79
100,000	31.97	31.94	33.03	35.92	37.92

Table 4. Processing time of the PIN^{hybrid} algorithm for different network sizes [s]

Nodes Edges	1,000	5,000	10,000	50,000	100,000
1,000	0.91	1.16	1.43	3.73	6.79
5,000	3.76	4.03	4.38	7.31	9.84
10,000	7.59	7.81	7.90	10.43	13.93]
50,000	35.77	35.51	35.67	38.90	43.84
100,000	69.44	70.57	71.50	76.87	80.01

network size on processing time is smaller than for the PIN^{nodes} algorithm. Moreover, this influence is comparable to the PIN^{edges} algorithm.

The comparison of different variants of the PIN algorithm reveals that the fastest one is always PIN^{edges}. See for example processing time for networks with the constant number of edges 50,000 and different number of nodes, Fig. 2. Note that in case of the PIN^{edges} and PIN^{hybrid} algorithms, processing times do not differ a lot among 1,000-, 5,000- and 10,000-node networks and they oscillate around 16 s for PIN^{edges} and 35 s for PIN^{hybrid}. The PIN^{edges} algorithm is 636.2 times faster than PIN^{nodes} for 1,000 nodes and 768.77 times faster for 100,000 nodes. Simultaneously, the PIN^{edges} algorithm is approximately two times faster than PIN^{hybrid} for all types of the investigated random networks with 50,000 edges, Tab. 5.

Processing time is a monotonic and increasing function of the number of nodes in the network, i.e. the greater number of nodes, the greater processing time, Fig. 2. However, only in case of the PIN^{edges} algorithm the processing time is

Table 5. The relation of processing times of PIN^{edges} to other PIN algorithms for the fixed number of edges (50,000)

No. of nodes	$\frac{t_{PIN^{nodes}}}{t_{PIN^{edges}}}$	$\frac{t_{PIN^{hybrid}}}{t_{PIN^{edges}}}$
1,000	636.20	2.16
5,000	675.38	2.19
10,000	685.97	2.18
50,000	775.65	2.05
100,000	768.77	1.92

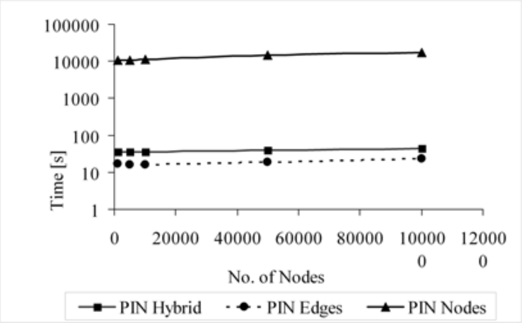


Fig. 2. Processing time in relation to the number of nodes for the fixed no. of edges (50,000)

Table 6. The ratio (tangent of slope angle) of processing time and number of nodes for different PIN algorithms, constant number of edges (50,000)

No. of nodes	PIN^{nodes}	PIN^{edges}	PIN^{hybrid}
1,000	10.5388	0.0016	2.1591
5,000	2.1876	0.0015	2.1924
10,000	1.1242	0.0015	2.1765
50,000	0.2940	0.0013	2.0523
100,000	0.1752	0.0013	1.9238

Table 7. The relation of processing times of PIN^{edges} to other PIN algorithms for the fixed number of nodes (50,000)

No. of edges	$\frac{t_{PIN^{nodes}}}{t_{PIN^{edges}}}$	$\frac{t_{PIN^{hybrid}}}{t_{PIN^{edges}}}$
1,000	39.07	1.05
5,000	170.91	1.45
10,000	329.68	1.59
50,000	775.65	2.05
100,000	749.28	2.14

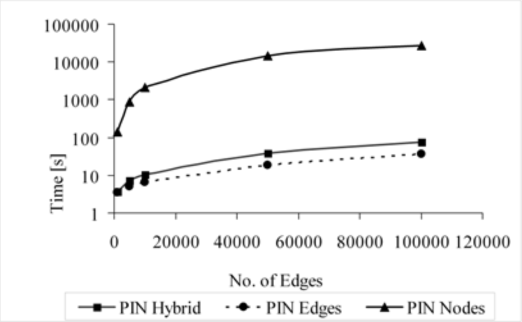


Fig. 3. Processing time depending on the number of edges, fixed no. of nodes (50,000)

a linear function of the number of nodes in the network. Moreover, the tangent of slope angle is very close to zero, i.e the values of the function increase very slow. In other words, they are almost constant, Tab. 6.

Let us consider the processing time for networks with the fixed number of nodes (50,000) but for the variable number of edges, Fig. 3. Note that, in contrary to networks with the constant number of edges, the processing times differ a

Table 8. The ratio (tangent of slope angle) of processing time and number of edges for different PIN algorithms, number of nodes equals 50,000

No. of edges	PIN^{nodes}	PIN^{edges}	PIN^{hybrid}
1,000	0.1388	0.0256	1.0514
5,000	0.1723	0.0059	1.4492
10,000	0.2161	0.0030	1.5905
50,000	0.2940	0.0013	2.0523
100,000	0.2692	0.0013	2.1397

lot among 1,000-, 5,000-, 10,000-, 50,000-, and 100,000-edge networks, for the PIN^{edges} and PIN^{hybrid} algorithms. It changes from 3.55 s for 1,000 edges to 35.92 s for 100,000 edges for PIN^{edges}, whereas for PIN^{hybrid}, it changes from 3.73 s for 1,000 edges to 76.87 s for 100,000 edges. The PIN^{edges} algorithm is 39.07 times faster than PIN^{nodes} for 1,000 nodes and 749.28 times faster for 100,000 nodes. Simultaneously, PIN^{edges} is as fast as PIN^{hybrid} for 1,000 nodes and two times faster for 100,000 nodes, Tab. 7.

Processing time is a monotonic and increasing function of the number of edges, i.e. the greater number of edges, the greater processing time, Fig. 3. However, the relationship cannot be seen as linear function of the number of edges in the network, Tab. 8.

5 Conclusions

Social position, which has been studied in this paper, is one of the measures useful to evaluate centrality of the node within the social network. Its iterative nature requires more or less iteration to be performed to achieve the required precision of results. However, the implementation of the general concept can be realized with different approaches. Three of them have been analyzed in the paper: PIN^{nodes}, PIN^{edges}, and PIN^{hybrid}. One of the most surprising conclusions from the tests carried out is the big difference in efficiency between these three methods, even over two orders of magnitudes. The ”edge approach” appears to be absolutely the best while raw, direct implementation of the concept – PIN^{nodes} remains far behind. This reveals that the implementation method for some general concepts from social network analysis may have the crucial impact on the computation efficiency. The future work will focus on the analysis of the effectiveness for other methods in social network analysis.

Acknowledgment

The work was supported by The Polish Ministry of Science and Higher Education, grant no. N N516 264935.

References

1. Alexander, C.N.: A method for processing sociometric data. Sociometry 26, 268–269 (1963)
2. Bavelas, A.: Communication patterns in task – oriented groups. Journal of the Acoustical Society of America 22, 271–282 (1950)
3. Brandes, U., Erlebach, T.: Network Analysis, Methodological Foundations. Springer, Heidelberg (2005)
4. Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C.: Wprowadzenie do algorytmó, Wydawnictwa Naukowo-Techniczne, Poland (2004)
5. Carrington, P., Scott, J., Wasserman, S.: Models and methods in Social Network Analysis. Cambrige University Press, Cambrige (2005)
6. Degenne, A., Forse, M.: Introducing social networks. SAGE Publications Ltd., London (1999)
7. Freeman, L.C.: A set of measures of centrality based on betweenness. Sociometry 40, 35–41 (1977)
8. Kazienko, P., Musial, K.: On Utilising Social Networks to Discover Representatives of Human Communities. International Journal of Intelligent Information and Database Systems, Special Issue on Knowledge Dynamics in Semantic Web and Social Networks 1(3/4), 293–310 (2007)
9. Kazienko, P., Musial, K., Zgrzywa, A.: Evaluation of Node Position Based on Email Communication. Control and Cybernetics 38 (1) (in press, 2009)
10. Kazienko, P., Musial, K.: Social position of Individuals in Virtual Social Networks. Journal of Mathematical Sociology (submittd, 2009)
11. Kendall, M.G.: Rank correlation methods. Charles Griffin & Company, Ltd., London (1948)
12. Proctor, C.H., Loomis, C.P.: Analysis of sociometric data. In: Jahoda, M., Deutch, M., Cok, S.W. (eds.) Research Methods in Social Relations, pp. 561–586. Dryden Press, NewYork (1951)
13. Sabidussi, G.: The centrality index of a graph. Psychmetrica 31(4) (1966)
14. Shaw, M.E.: Group structure and the behavior of individuals in small groups. Journal of Psychology 38, 139–149 (1954)
15. Wasserman, S., Faust, K.: Social network analysis: Methods and applications. Cambridge University Press, New York (1994)

Time Series Analysis of R&D Team Using Patent Information

Yurie Iino[1] and Sachio Hirokawa[2]

[1] Graduate School of Information Science and Electrical Engineering,
Kyushu University
akioiino@dion.ne.jp
[2] Research Institute for Information Technology, Kyushu University
hirokawa@cc.kyushu-u.ac.jp
http://matu.cc.kyushu-u.ac.jp

Abstract. Reliable real data is indispensable for the examination, evaluation and the improvement of the organizational structure. This paper proposes a method to use patent documents for analyzing organizational structure of researchers. The method is more efficient and objective compared to personal interview. The structure of research groups is modeled as a "inventors graph", which is a directed graph where each node represents an inventor and an edge represents co-inventor relationship. Empirical evaluation is conducted to cosmetic related companies and their patents that applied between 1998 and 2002 in Japan. It is shown that there is different characteristics in the inventors graph between Japanese companies and foreign companies. Moreover, time series analysis revealed that the inventors graphs of a Japanese company Kao changed in 2001 to foreign company type.

1 Introduction

It is important for enterprises to know the technological development activity of competitors so as to evaluate strong point and weakness of own company for deciding policy of future business promotion. Patent documents are reliable resource for analyzing competitors. Patents are considered as results of investment in research activities and guarantee the monopoly of technology and the exclusion of other's making to right. Application tendency reveals the company's evaluation and judgement for technological fields. Indeed, a strong relationship is reported between R&D investment and patent application trend [5].

Technological development requires many researchers. Connection of researchers creates the flows of new technologies. The relation of inventors becomes a clue to know the relation of technological development [1]. In [3,4], we showed a method to construct network of inventors using patent information. The present paper applies the method for cosmetic related companies and expands the method for time series analysis of R&D structure.

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 464–471, 2009.

2 Analysys of R&D Structure Using the Number of Patents and Inventors

We used 14,857 patents for analysis. They are applied in Japan from 1998 to 2002 and are cosmetic related patents classified by the IPC A61K7. The patent data have been retrieved through the commercial patent service JP-Net [1]. No preprocessing is done for variation of names of company and people. Table 1 shows the basic statistics of data; the number of applications, the number of inventors and the average number of patents by an inventor. The table compares the top three Japanese companies and foreign companies. Foreign companies, except L'Oreal, have half of average patents compared with the Japanese companies.

Table 1. Patent Activities of top cosmetic companies

Company	#Applications(AP)	#Inventors(IV)	AP/IV
KAO	997	709	1.4
SHISEIDO	1025	504	2.0
LION	760	456	1.7
L'OREAL	1089	842	1.3
P&G	383	599	*0.6
UNILEVER	238	506	*0.5

The simple analysis in Table 1 reveals a difference between foreign companies and Japanese companies in average number of patents by an inventor. However, we cannot understand how inventors are related and how research groups are formed for a particular company or for a specific target field.

In the later sections, we analyse not an individual inventor, nor an average inventors but the relationship of inventors and how the research groups are organized in particular companies.

3 Inventors Graph

The notion of concept graph is introduced in [10] to represent relationship between keywords that appear in the search result documents. An ordered edge of a concept graph represents hypernym/hyponym relation of keywords.

The order relation of words are formulated as follows. A word w is a *characteristic* word of the search result $D(q)$ of a query q with respect to a threshold α if the word w satisfies the condition $df(w, D(q))/df(w, U) > \alpha$. For characteristic words u and v, u is said to be *hypernym* when both of the conditions $df(u, U) > df(v, U)$ and $df(u * v, D(q))/df(v, D(q)) > \alpha$ hold. Here, U is the set of all documents being considered, $df(w, X)$ represents the number of documents in $X \subseteq U$ that contain w, $df(u * v, X)$ represents the number of documents in X that contain both u and v, and α is called a threshold.

[1] http://www.jpds.co.jp

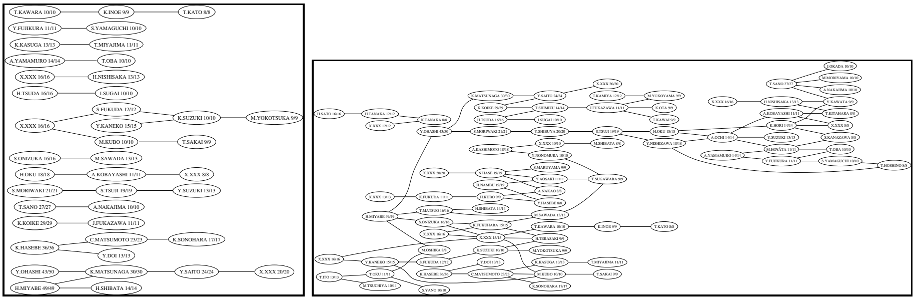


Fig. 1. Kao Inventors Graphs with $\alpha = 0.7$(left) and $\alpha = 0.1$(right)

Note that the set of characteristic words and the hypernym/hyponym relation is determined with respect to the target document set. The concept graph of $D(q)$ is a directed graph whose nodes are characteristic words of $D(q)$ and whose edges are the hypernym/hyponym relation between keywords.

Given a query, the concept graph visualises the whole picture of the search results and gives hints for further search and for comprehension.

In the present paper, the company name is used as a query, and the names of inventors are considered as keywords of concept graph. The graph represents the relationship of researchers of the company. Inventors that form a connected compoment in a concept graph belong to the same research group and key inventors appear in the uuper position of the connected component.

The researchers who invented a large number of patents are shown on the left side of an edge, while the researchers on the right side of an edge has relatively small number of patents. Roughly speaking, researchers on the left side can be considered in higher level, and the researchers on the right side are in lower level. When the ratio is larger than the threshold, a higher researcher is connected with a lower researcher with an edge.

In the concept graph of the threshold $\alpha = 0.5$, more than half of the patents by a right side inventor are co-invented by the left side inventor. When the threshold $\alpha = 0.1$, more than one tenth are co-invented with the higher level researcher. Therefore, lower the threshold α, more the weak relationships are extracted and displayed in the corresponding concept graph.

Fig. 1 shows the relationship of inventors of Kao, who invented more than 8 patents, with threshold $\alpha = 0.7$ and $\alpha = 0.1$. We see several independent groups in the graph with $\alpha = 0.7$. On the other hand, in the graph of $\alpha = 0.1$, these separated groups are complicatedly linked each other by inventors of lower level and form a large group.

4 Inventors Graph of Japanese Companies and Foreign Companies

There is a clear difference in the average number of patents by an inventor between three Japanese companies and two foreign companies, as we see in

Table 1. However, the average number for the foreign company L'Oreal is 1.3 which is almost same to that of Japanese companies.

Reconsider the inventors graphs of Kao with $\alpha = 0.7$ and $\alpha = 0.1$ (Fig. 1). The isolated groups in the graph with $\alpha = 0.7$ are connected in that with $\alpha = 0.1$. According to [11], Kao adapts the "matrix" organization to form a R&D team where researchers with different background knowledge are merged to generate interactive effects. The nested graph of inventors obtained by a lower threshold agrees to this matrix organization. We observe the similar pattern in other Japanese companies. This implies that there is a diversity of strength of relationship between researchers in Japanese companies.

We can think of the following reasons that would explain the change of the concept graphs for Kao with different threshold α.

- A large number of researchers engage in developing a common basic technology. However, in individual product development, researchers form divided groups with a few people and flexibly cooperate between them.
- In the development of the core technology, several different sections work together to realize effective right as fast as they can.
- In a large project, there are many sections working together.

To compare Kao and L'Oreal , we draw the concept graphs of inventors for the two companies in Fig.2 with the threshold $\alpha = 0.1$.

The relationship of researchers in L'Oreal are tighter than that of other Japanese companies. Very small number of researchers are connected to each other. We can guess that they do not adapt the flexible formation of research groups.

This observation coincides with the report [9] of difference of R&D management between French companies and Japanese companies.

According to [9], the role and the commission of a section are fixed in French companies and researchers prefer to move out of the company for their

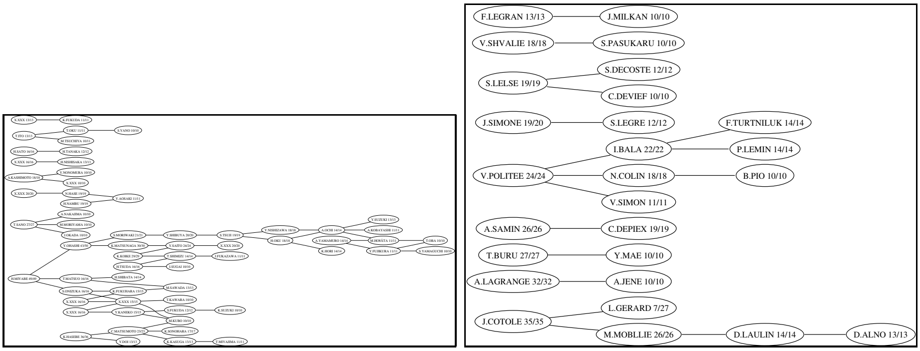


Fig. 2. Inventors Graph of Kao(left) and L'Oreal(right)

promotion. On the other hand, in most Japanese firms who adpt life-time employment system, researchers stay in a company and move in and out from R&D division to other sections such as production lines.

5 Time Series Analysis of R&D Structure of a Japanese Company

Evaluation and reorganization are necessary to improve the efficiency of research and development. The analysis of the previous section was obtained from the whole data in 5 years. Kao formed a very large group which consists of several medium and small sub-groups linked with variety of strength. In this section, we compare the inventors graphs of Kao by each year and consider the structural change of the company.

Table 2 shows the number of co-inventors for a patent in each year. There are small increases and decreases in the total number of patents and inventors in every year. The average number of patents by one researcher has no big change. But, the average number of co-inventors of a patent has the peak in 1999. After that year, the number of co-inventors declined and the number of patents by a single inventor increased. We can guess that there seems to be some tendency in decline of the linkage strength. However, we cannot understand how actual relationship were changed.

To make an detailed analysis of relationships of inventors, we draw Fig. 3 for each fiscal year, where we restricted to the researchers who have more than 2 patents in the year. The inventors graph for the patents of Kao applied in 1998 has the similar characteristics that we mentioned in the previous section, where the changed of the threshold α brings the change of the structure of the graph. However, the inventors graphs for 2001 and later does not change when the threshold is changed. In other words, the graphs for 2001 and later became similar to those of foreign companies. This suggests a hypothesis that there should be some organizational change in 2001 for Kao company. Indeed, the company introduced the management indicator EVA^{TM}(Economic Value Added) in 1999, and in June 2000 the indicator was applied for personnel assessment of the whole employee [2].

Table 2. The number of co-inventors

	#co-inventors								total	ratio of	average
year	1	2	3	4	5	6	7	8	#patents	single inventor	#co-inventors
1998	20	66	58	32	24	12	4	1	217	0.09	3.1
1999	15	48	49	35	25	12	2	0	186	0.08	3.3
2000	30	66	57	27	30	6	0	2	217	0.14	2.9
2001	31	55	47	19	9	1	0	1	163	0.19	2.6
2002	56	78	43	13	14	10	0	0	214	0.26	2.4

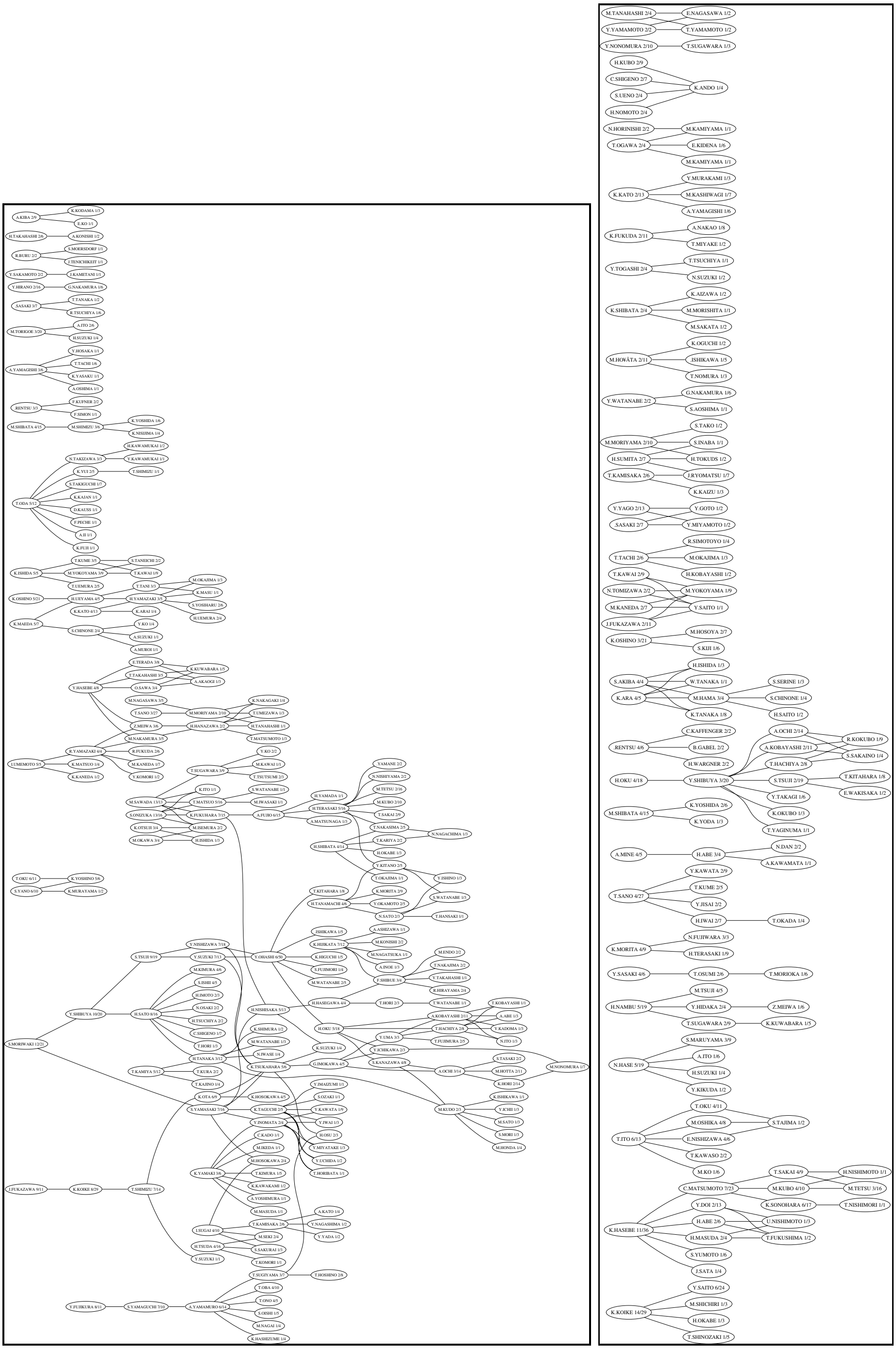


Fig. 3. Kao Inventors Graph of 1998(left) and 2001(right)

6 Related Work

There are two aspects in analyzing network of people – ego-centric and socio-centric. Ego-centric analysis focuses on key person and relation to other people. Socio-centric analysis consider all relationship between each member of the target community. When we analyse inventors of a company, the result would be different according to the point of view. In the former case, individual researchers are the target of the analysis for extracting top inventors of the company. In the later case, to overview the whole picture of R&D activities is the aim of the analysis. The inventors graph shown in the present paper can be used in both aspects of ego-centric and socio-centric analysis.

In [1], Breitzman showed examples of "co-inventor brainmap" of the target enterprises, where key inventors are listed on the vertical axis and the patent IDs are listed on the horizontal axis according to application date. Co-inventors of the same patent are plotted on the vertical line that corresponds to the co-invented patent. Thus, a brainmap displays one-to-one correspondence of inventors and chronological change of key inventors.

A tree chart is used to display inventor correlation in [6]. The purpose of the system is to grasp the application trends of the key inventors who are drawn in the roots of the tree.

Sugiyama et.al. [12] used a network of inventors to extract the core company in allied companies where they analyse how technological innovation are formed and knowledge are transferred.

Nakai [7] applied the similar method to draw network of co-inventors where key words are attached on edges. They claim that the tool is useful to investigate the process of technological innovation. In [8], they confirmed the effectiveness of the method by applying to patent data and to scientific articles, where they found characteristics of basic research groups and that of industrial research groups and found how the technologies have been developed.

All of those methods visualise R&D team structure of target companies in particular period. But as far as the authors know, there is no previous research that used those methods for time series analysis of organizational change of R&D team structure.

7 Conclusion

This paper showed analysis of R&D structures of cosmetic related companies using inventors graph where a node represents an inventor and an edge represents co-inventor relationship of researchers. It is shown that Japanese companies and foreign companies have different characteristics in their graphs. By time series analysis of Kao company, it turned out that the inventors graph changed in 2001 when the company introduced the management indicator EVA^{TM} for personnel assessment. The propose method is easy to apply and effective to understand the group structure of R&D team compared to conventional sociometric method such as interview.

References

1. Breitzman, M.: The many applications of patent analysis. J. Information Sci. 28(3), 187–205 (2002)
2. Fukuda, T.: EVA Management System in Kao Corporation(in Japanese). Quarterly journal of economics, Economic Research Institute of Kanto Gakuin University 208, 12–26 (2001)
3. Iino, Y., Yamada, Y., Hirokawa, S.: Structural Analysis of R & D Division from Patent Documents. In: Proc. IEEE International Conference on e-Business and Engineering, pp. 423–428 (2008)
4. Iino, Y., Hirokawa, S.: Analysis of R & D system of cosmetic companies based on patent information (in Japanese). In: Proc. 4th International Symposium Technological Innovations in Japan, pp. 43–48 (2008)
5. Japan Patent Office, Distributed Material No.1 (in Japanese). In: Industrial Property Council International Division 5th meeting (1999)
6. Miyake, M.: Information Analysis Method, Program, Recorded Medium and Apparatus, Japan Patent Application No. 2005-277349 (2005)
7. Nakai, T.: Patent portfolio analysis by text mining (in Japanese). Journal of Information Processing and Management 51(3), 194–206 (2008)
8. Nakai, T., Sakauchi, S., Yamaguchi, Y., Nakatani, I.: Information Visualization by Integration of Patent Data and Scientific Articles. In: Proc. 5th INFOPRO symposium, pp. 81–85 (2008)
9. Setani, M., Arico, F., Lambert, G., Llerena, P.: A comparative study of R&D management between Japanese and French Companies (in Japanese). National Institute of Science and Technology Policy, Japan (1997)
10. Shimoji, Y., Wada, T., Hirokawa, S.: Dynamic Thesaurus Construction from English-Japanese Dictionary. In: Proc. The Second International Conference on Complex, Intelligent and Software Intensive Systems, pp. 918–923 (2008)
11. Shiraishi, H.: Constitution that promotes research and development (in Japanese). Kanazawa University Department of Economics Bulletin Paper 27(1), 171–231 (2007)
12. Sugiyama, Y., Takao, Y., Ku, S., Kubo, R.: Management of Ecosystem Formation and Knowledge Transfer by Core Laboratory in "Keiretsu" Network, Technical Report of Industrial Technology Research Grant Program in FY2006 (in Japanese), New Energy and Industrial Technology Development Organization, Japan (2008), http://www.tech.nedo.go.jp/PDF/100011314.pdf

Extracting Research Communities by Improved Maximum Flow Algorithm

Toshihiko Horiike[1], Youhei Takahashi[1], Tetsuji Kuboyama[2],
and Hiroshi Sakamoto[1]

[1] Kyushu Institute of Technology, Kawazu 680-4, Iizuka 820-8502, Japan
{t_horiike,y_takahashi,hiroshi}@donald.ai.kyutech.ac.jp
[2] Gakushuin University, 1-5-1 Mejiro Toshima Tokyo, 171-8588, Japan
kuboyama@tk.cc.gakushuin.ac.jp

Abstract. In this paper we propose an algorithm, which is an improvement of identification of web communities by [1], to extract research communities from bibliography data. Web graph is huge graph structure consisting nodes and edges, which represent web pages and hyperlinks. An web community is considered to be a set of web pages holding a common topic, in other words, it is a dense subgraph of web graph. Such subgraphs obtained by the max-flow algorithm [1] are called *max-flow communities*. We then improve this algorithm by introducing the strategy for selection of community nodes. The effectiveness of our improvement is shown by experiments on finding research communities from CiteSeer bibliography data.

1 Introduction

In this paper we propose an improvement for the maximum flow algorithm to find dense subgraphs as web communities. An web community is a set of web pages holding a common topic, which is represented by a connected subgraph in web graph. Study of extracting web community has attracted many researchers since its wide application to web technology, like trend discovery and information recommendation.

In the last decade many methods for extracting web communities were presented. In [1,2,3], algorithms aim to find dense subgraphs using local information of web graph, and in [4,5,6], algorithms extract communities by using global information, like HITS [7]. The problem handled in this paper is to extract research communities from bibliography data, and we focus on the former research strategy, i.e. extracting dense subgraphs using local information. This problem motivate us to extract interesting research communities as follows.

A simpler idea for dense subgraphs is to find large cliques in the graph, where a clique of a graph is a complete subgraph in it. However such the problem of finding large cliques is computationally hard, so several other methods were presented. One of such alternative methods is to find bipartite graphs and the other is to find dense subgraphs defined by max-flow problem in networks.

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 472–479, 2009.

In [6], a community is defined as a subgraph which contains at least one clique, and all communities in input web graph are enumerated. In this method, indegree/outdegree of nodes are closely related to extracted communities. Consequently a small degree node is hardly selected as a member of a community even if it is an important node. This method is basically equal to the notion of hub-authority in [7].

Such a method based on hub-authority is effective for extracting a global relation in a graph. However, not all important communities are extracted by this method. For instance, consider the problem to find communities from research data, which is a bibliography consisting all related references. An important community is constructed by several pioneering studies and other related studies. The extraction of such communities by the hub-authority method is difficult, since there is very few study which impacts on the whole research field or many different research fields. In order to extract such *compact* relations, we adopt the strategy of maximum flow community.

In [1,2], the definition of an web community that enables web crawlers to effectively focus on narrow but topically related subsets of web and also enable search engines and portals to increase precision and recall of search results. They define a community to be a set of web pages that link to more web pages in the community than to pages outside of the community. Generally the hardness of extracting such communities depends on a *priori* information presented for algorithms. In fact, in the absence of any priori information, the problem becomes to be NP-hard due to graph partitioning problem. However by exploiting various properties of web, identifying web communities becomes identical to solving the maximum flow network problem, which is solvable in polynomial time in the size of input network.

A polynomial time algorithm for max-flow network problem is presented in [8]. Flake et al. introduced an efficient method of extracting communities using such max-flow algorithm. So we call a community extracted by the max-flow algorithm a *max-flow community.*

Efficiency of max-flow communities depends on the ranking of community nodes extracted by algorithm. In this paper we propose a modification for the ranking by careful evaluation according to connectivity to seed nodes. In original strategy, ranking is decided by only the number of sum of indegree and outdegree of an extracted node. Such communities are too sensitive to the number of edges, and there is a possibility that irrelevant nodes are associated to a community due to their many edges. We thus claim that ranking of community nodes should be evaluated by also the relation to seed nodes, which are some cores of a community, i.e. some nodes deeply related to seeds should be ranked as higher positions.

In this paper we implement the modified max-flow community algorithm and show its efficiency by experiments. Particularly this algorithm is applied to CiteSeer bibliography data and many interesting communities are obtained from a small number of seed nodes.

The remainder of this paper is divided into four additional sections. In Section 2 we give an overview of max-flow community algorithm. The definitions and notations are given in this section. Section 3 contains the idea for modification of max-flow algorithm. In our method, ranking is computed according to the connectivity between extracted nodes and seed nodes as well as the degree of nodes. In Section 4 we present the experimental results in CiteSeer bibliography data. Compared to web structure, the graph represented by bibliography data hardly contains large communities since there is very few study which impacts on whole research area. So we need other method to extract smaller communities. Our modification intends to extract such special communities, and the experimental results shows the effects. Finally we discuss a future work on this study.

2 Max-Flow Community

In this section we summarize the max-flow algorithm by [8] and the web community extraction algorithm by [1].

Maximum Flow and Minimum Cuts

The $s-t$ maximum flow problem is defined as follows: Given a directed graph $G = (V, E)$, with edge capacities $c(u, v)$ of positive integers, and two nodes $s, t \in V$, find the maximum flow that can be routed from *source* s to *sink* t that obeys all capacity constraints. Intuitively, if edges are water pipes and nodes are pipe junctions, then the maximum flow problem tells us how much water we can move from s to t.

The max-flow min-cut theorem in [9] proves that $s-t$ maximum flow of a network is identical to *minimum cut* for separating s and t. Many polynomial time algorithms exist for solving the $s-t$ maximum flow problem, and applications of the problem include VLSI design, routing, scheduling, image segmentation, and network reliability testing.

The maximum flow problem is well-suited to the application of identifying web communities since it is computationally tractable and it allows us to exploit a *priori* knowledge about underlying graph.

Most modern solutions to the max-flow problem operate under the assumption that entire graph under consideration can be examined easily. This is obviously not the case with web, as entire graph that corresponds to web is vastly larger than any single computer can store in main memory. Nevertheless, one of the simplest max-flow algorithms [10] can solve the problem by examining only portions of graph that arise when locating shortest paths between s and t. Thus, it should be possible to solve the max-flow network problem on entire web.

Web Communities

We next formalize web communities obtained by max-flow solution for web graph.

Definition 1. (G.W. Flake et al. [1]) A community of undirected graph $G = (V, E)$ is a subset $C \subseteq V$ such that for all nodes $v \in C$, v has at least as many edges connecting to nodes in C as it does to nodes in $(V - C)$.

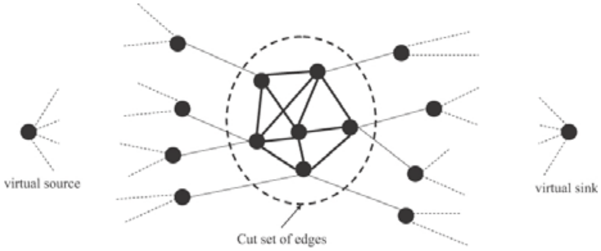


Fig. 1. An example of the minimum cut and the web community

It is easy to expand this definition to directed graphs. In the next section we propose the improved community algorithm. Here we just summarize the original algorithm by [1]. In web graph, each node is corresponding to an web page, and an edge (u, v) is a link from u to v with a unit capacity $c(u, v) = 1$. Initially we assume a set S of seed nodes and a set $C = \emptyset$ of community nodes.

For any nodes s, t, we can compute $s - t$ maximum flow such that s, t are separated by a minimum cut set of saturated edges. Remove all such saturated edges, and all nodes reachable from a node in S are added to C. Next we compute the ranking of nodes in C with respect to their indegree/outdegree numbers. Upper ranked nodes are moved to S, and we continue the above process until C is steady. Then obtained connected subgraphs are called max-flow communities.

In [1], we assume a virtual source s and a virtual sink t with $c(s, v) = \infty$ and $c(v, t) = 1$ for any $v \in V$. They proved that such web communities with virtual source/sink are extracted by this algorithm without a *priori* knowledge about web graph.

In Fig. 1, we give an example of max-flow community. The virtual source/sink is assumed outside of the web graph. For the computed max-flow, we obtain the corresponding min-cut illustrated by the circle of broken line.

3 Improving Max-Flow Algorithm

In this section we explain our algorithm based on the max-flow community algorithm by [1]. For an web graph, given a set of seed nodes, S, if an iteration of the algorithm, we obtain a set of community nodes, C. For the next iteration, we must compute the ranking of all member of C, which is defined by the number of degrees. In this setting, the ranking of a community node $v \in C$ is defined by the value $indegree(v) + outdegree(v)$, which is independent of the number of seed nodes directly associated with v, while the relation between S and C is important information for our communities. We thus propose an idea that the ranking of C should be computed with respect to the number of edges associated with S, i.e. $indegree(v, S) + outdegree(v, S)$.

In Fig. 2, we illustrate this ranking method. The seed nodes are denoted by the black nodes, and the community nodes are labeled by integers 1, 2, 3, and 4. Since node 1 is maximum in the measure $indegree(v) + outdegree(v)$, this node

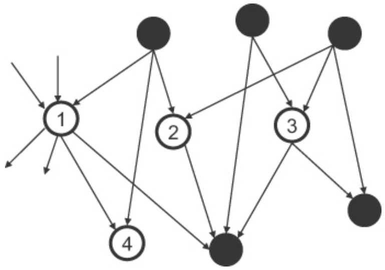


Fig. 2. Seed nodes and ranking of community nodes

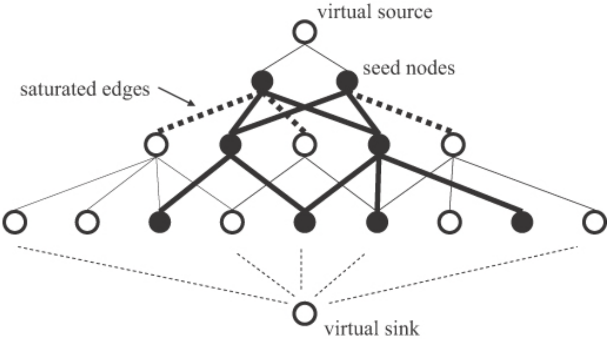


Fig. 3. The flow of community extraction

is ranked in the top by the original method. On the other hand, in our ranking method by $indegree(v, S) + outdegree(v, S)$, node 3 is ranked in the top, and the community nodes are sorted as $3, 2, 1, 4$. Using the new ranking measure, we describe the improved max-flow community algorithm below.

Improved Max-Flow Community Algorithm

Input: The set of seed nodes, S, and an web graph $G(V, E)$ whose nodes are reachable from a seed node within 2 edges.

Output: Web community C.

Preprocess: For $(u, v) \in E$, set the capacity $c(u, v) = |S|$, for the virtual source s and sink t, and $v \in V$, set $c(s, v) = \infty$ and $c(v, t) = 1$, and let $C = \emptyset$.

(1) Execute the max-flow algorithm on G.

(2) Compute all $v \in V$ which are reachable from an $s \in S$ by only unsaturated edges, and add all of them to C.

(3) Decide the ranking for $c \in C$ by $indegree(c, S) + outdegree(c, S)$, and move the higher ranked nodes from C to S.

(4) Continue (1)-(3) until C become to be steady, and output C.

In Fig. 3, we show flow of extracting communities from web graph. The highest node is the virtual source and the lowest is the virtual sink. Any seed node is associated by the source, and the sink is associated by any node. By the step (1) of the algorithm, a maximum flow is obtained. At this time, the saturated edges, which are illustrated by bold broken lines in Fig. 3, denote the cut edges. Intuitively, an extracted web community is consisting of nodes which are reachable from the source without cut edges. In the next section we examine the efficiency of our improvement.

4 Community Extraction from Bibliography Data

In this section we show experimental results on max-flow community extraction from CiteSeer bibliography data [11], which includes over 700,000 entries for research documents. The following example is a typical record in this data, where information of author and abstract is removed beforehand.

```
<record>
<id>7348</id>
<title>Parallel Sorting by Overpartitioning</title>
<ref>14421,40374,91922,40140,372786,4945,8848</ref>
</record>
```

In this data, any paper has its unique ID defined inside `<id>` tag. In the above example, paper `7348` is referring other studies indicated by 7 integers inside `<ref>` tag. Thus, we can regard a collection of such records as a directed graph.

For such a bibliography graph, we examine the efficiency of the modified max-flow community algorithm compared to the original one. In this experiment, we choose “sort” as a keyword, and select seed nodes which contains this keyword in title or abstract. Fig. 4 shows the result of the precision of our proposed method compared to that of the standard max-flow community algorithm [1] for different seed nodes related to “sort”.

Here the precision of a community in Fig. 4 is defined by

$$\text{precision} = \frac{\text{\# of records related to the keyword in the community}}{\text{\# of records in the community}} \times 100,$$

where *record related to the keyword* means ’record containing the keyword in its title or abstract’. By this experiment, we can obtain higher precision compared to the standard max-flow community algorithm even for different seed nodes.

Finally we illustrate a sample of communities extracted by [1] and our algorithm in Fig. 5. These two communities are obtained by selecting seed node 7348. In this figure, the nodes with underlined numbers are irrelevant to the keyword. By the results, we conclude that compact communities are obtained from bibliography data by our improved community extraction algorithm.

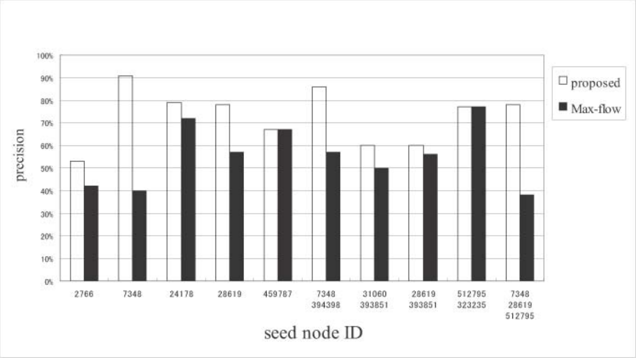


Fig. 4. Comparison between the max-flow and proposed community

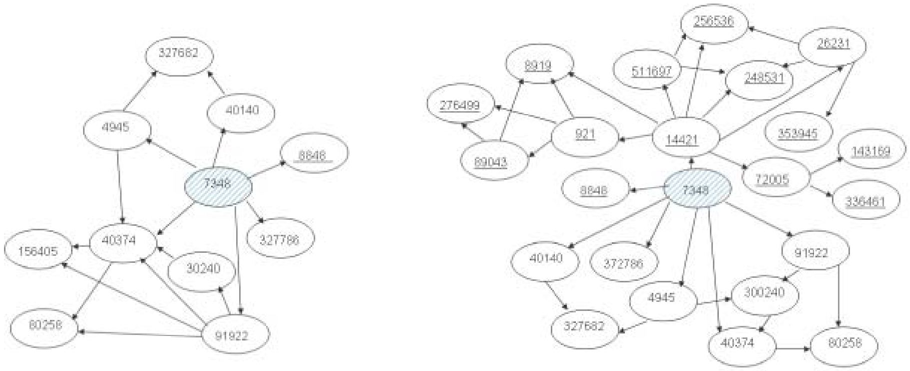


Fig. 5. Comparison of two communities: the left one by the improved algorithm and the right one by the standard max-flow method

5 Conclusion

In this paper we propose an improvement of the ranking of community nodes by carefully evaluating the relation between seed nodes and community nodes. The effectiveness of our improvement is shown by experiments for finding research communities from CiteSeer bibliography data. More compact and close communities are obtained by our algorithm compared to the standard max-flow community algorithm by [1]. We thus conclude that our algorithm is effective for extracting relatively compact communities from directed graph data.

As future work we would develop a hybrid algorithm for community extraction. In [12], comparison of two types of algorithms based on complete bipartite graphs and max-flow network was presented, and it was reported that more generic communities are obtained by the former method and more specific communities are obtained by the latter method. In [13], a method for extracting relation among web communities using HITS was proposed. We thus try to expand our strategy for extracting compact communities to the above different types of community extraction.

References

1. Flake, G.W., Lawrence, S., Giles, C.L.: Efficient identification of web communities. In: KDD 2000, pp. 150–160 (2000)
2. Flake, G.W., Lawrence, S., Giles, C.L., Coetzee, F.: Self-organization and identification of web communities. IEEE Computer 35(3), 66–71 (2002)
3. Kumar, R., Raghavan, P., Rajagopalan, S., Tomkins, A.: Trawling the Web for Emerging Cyber-Communities. Computer Networks 31(11-16), 1481–1493 (1999)
4. Chakrabarti, S., Dom, B., Raghavan, P., Rajagopalan, S., Gibson, D., Kleinberg, J.M.: Automatic resource compilation by analyzing hyperlink structure and associated text. Computer Networks 30(1-7), 65–74 (1998)
5. Gibson, D., Kleinberg, J.M., Raghavan, P.: Inferring web communities from link topology. In: Hypertext 1998, pp. 225–234 (1998)
6. Kumar, R., Raghavan, P., Rajagopalan, S., Tomkins, A.: Extracting Large-Scale Knowledge Bases from the Web. In: VLDB 1999, pp. 639–650 (1999)
7. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. In: SODA1998, pp. 668–677 (1998)
8. Goldberg, A.V., Tarjan, R.E.: A new approach to the maximal flow problem. In: STOC 1986, pp. 136–146 (1986)
9. Ford Jr., L., Fulkerson, D.: Maximal flow through a network. Canadian Journal of Mathematics 8, 399–404 (1956)
10. Edmonds, J., Karp, R.M.: Theoretical improvements in algorithmic efficiency for network flow problems. J. ACM 19(2), 248–264 (1972)
11. CiteSeer.IST, http://citeseer.ist.psu.edu/
12. Imafuji, N., Kitsuregawa, M.: Effects of maximum flow algorithm on identifying web community. In: WIDM 2002, pp. 43–48 (2002)
13. Toyoda, M., Kitsuregawa, M.: Creating a Web community chart for navigating related communities. In: Hypertext 2001, pp. 103–112 (2001)
14. Imafuji, N., Kitsuregawa, M.: Finding a web community by maximum flow algorithm with hits score based capacity. In: DASFAA, pp. 101–106 (2003)
15. Dean, J., Henzinger, M.R.: Finding Related Pages in the World Wide Web. Computer Networks 31(11-16), 1467–1479 (1999)
16. Asano, Y., Nishizeki, T., Toyoda, M., Kitsuregawa, M.: Mining communities on the web using a max-flow and a site-oriented framework. IEICE Transactions 89-D(10), 2606–2615 (2006)

Virtual Communities of Practice's Purpose Evolution Analysis Using a Concept-Based Mining Approach

Sebastián A. Ríos[1], Felipe Aguilera[2], and Luis A. Guerrero[2]

[1] Department of Industial Engineer, University of Chile
srios@dii.uchile.cl
[2] Department of Computer Science, University of Chile
{faguiler,luguerre}@dcc.uchile.cl

Abstract. Today, social networks systems have become more and more important. People have change their way to relate and communicate. Therefore, how to enhance contents and organization of a social network is a very important task. This way, we can help Virtual communities of practice (VCoP) to survive through time. VCoP are special kind of social network where the purpose is a key aspect. However, administrators are blind when trying to identify how to enhance the community. We propose a method which helps them by analyzing how purpose evolves through time. The approach has been experimentally tested in a real site with successful results.

1 Introduction

The WWW, has become a fertile land where anyone can transform his ideas into real applications to create new amazing services. Therefore, it was just a matter of time until the massive proliferation of virtual communities, social networks, etc. New social structures have been formed by massive use of new technologies. This way, people can relate to other by interests, experiences or needs.

In a scenario where WWW has become more important every day, and people is using more often the web to relate to others, to read news, obtain tickets, etc. The need to well organized web site has become one of the vital goals of enterprises and organizations. In order to accomplish such task web mining area was born about a decade ago.

Web mining are techniques to help managers (or experts) to extract information from a web site content, link structure or visitors' browsing behavior. This way, it is possible to enhance a web site, obtain visitors' interests patterns to create new services, or provide very specific adds depending on interests of visitors ([4,5]).

Today, virtual communities have experienced an exponential growth. Also, the use of web mining techniques to explode data stored in these systems has become a natural approach to obtain knowledge from them. However, a virtual community is not only a group of people accessing a web site, they establish social

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 480–489, 2009.

relationship through the use of Internet tools ([12]), allowing the formation of a communal identity and a shared sense of the world ([13]). In order to provide truly valuable information to help managers or web masters it is necessary to take into account the social nature of the virtual communities in web mining techniques.

Although, studies of social aspects in a virtual community are crucial, the evolution analysis of such aspects is more important yet, especially in social structures that allow to its members to define themselves different levels of participation among time, like Virtual Communities of Practice (VCoP) ([14]), provoking that community change according to member's participation.

This work focuses on the application of a novel web usage mining approach to study the social nature aspects of Virtual Communities. Specifically, we focus our work on how VCoP accomplish its purpose through the evaluation of the evolution of its goals' achievement.

2 Related Work

All virtual communities are different because of their social nature, however, there exist several common characteristics which allow to classify them ([11]). We can find interest communities, purpose communities and practice communities.

The VCoP's studies are yet immature, in spite of the studies already developed to analyze interest and purpose communities. For example, in the case of interest communities it is frequent to find recommender systems, evaluation/review systems, ranking systems, etc. Therefore, present paper is focused on VCoP.

2.1 Virtual Communities of Practice

Virtual communities of practice (VCoP) are informal, self-organizing networks of people dedicated to sharing knowledge ([14]). An important characteristic of VCoP is that very commonly some members help other to answer questions and solve problems. This way, they share and create knowledge (shared information, good practices, generate software tools and knowledge bases) ([14]).

Virtual communities of practice evolve in time passing through diverse levels of maturity ([14]). Through time, a VCoP can discover (objectives, needs, etc.) and re-invent itself ([6,14]). Therefore, to be able of study a VCoP it is not enough to perform analysis in a specific time period. It is of major importance to develop new algorithms or techniques to analyze how the VCoP evolve trough time. This way, managers or administrators will be able to discover risk situations or how the users interests are changing.

2.2 VCoP Analysis Techniques

Based on ([7]) we can identify four ways to analyze a VCoP which are commonly used. However, most of these approaches don't consider evolutionary factors in the analysis.

1. Ethnography and associated techniques: The purpose of ethnographic research is to build a rich understanding of a group or situation from the point of view of its members/participants. Disadvantage of this kind of analysis is that *in situ* studies are intrusive.
2. Questionnaires: This kind of analysis are useful for collecting demographic information and have the advantage that they can be distributed by hand to local participants, or posted via email or on the Web ([6,7]). Although, questionnaires provides useful information from virtual community members ([1]), they are not sufficient. It is recommended the use of a secondary technique, when possible, to reduce the subjectivity of members' judgements.
3. Experiments and quasi-experiments: Laboratory studies are valuable for testing the usability of the interface and users' reactions to new user interface features. To apply this approach it is needed to create the virtual community in a laboratory. This is quite complex since it is needed to have a representative group of virtual communities members, in a real (physical) experiment.
4. Data Mining & Social Network Analysis (SNA): This kind of analysis consist of using software's logs to discover useful information from the VCoP. This way, the main point is the study of communities' nature, i.e. the study of its members and the relations they establish ([6]). However, SNA analysis is designed to answer questions like: Who is the expert or experts in the community? Which sub communities or subgroups exists? But, it is not possible to answer: which is the interests of members? Which is the real goal of the community? Which topics are related with the community purpose? Which themes diverge from communities' main purpose? All these question remain unanswered.

We propose that SNA techniques are not sufficient to fully understand a virtual community of practice. Therefore, in this paper we propose a novel method to analyze a VCoP, based on a data mining approach to discover useful information from the VCoP for its better administration.

3 Proposed Methodology

The proposed methodology is based in the VCoP's purpose study. The purpose refers to a community's shared focus on an information need, interest, service. Every user has his own motivations (or interests) to use a specific VCoP and of course every community of practice can satisfy specific users' needs.

Defining the community's purpose is of major importance, since potential participants can immediately find out about the communities goals ([6]). However, VCoP's purpose is not always clear, even worst, it is not clear if all community members are aligned to the same common purpose.

But also, we must consider the evolutionary nature of purpose. Since, people change their motivations, interests, taste, etc. every day. Certainly, members of a VCoP change their purpose when using the VCoP. As example, the purpose of a newcomer are to learn the basics of a theme using the VCoP. When this newcomer becomes an expert his purpose might be to research specific and complex aspects of such theme; or even more, his purpose might be to answer questions of new comers.

Therefore, our hypothesis is that study of evolutionary nature of purpose of VCoP's members is a vital way to understand a community and to enhance it. Until today there aren't significant studies to evaluate this evolution (as we have shown in section 2), and fulfillment of purpose through time. Also, we strongly support the idea that the study of social networks members' relations alone (SNA) it is not sufficient to obtain a good understanding of a VCoP. It is needed to analyze other social aspect, such as purpose, to fully understand and perform the necessary enhancements to allow VCoP exists through time.

3.1 Goals as a Purpose Accomplishment Measure

Of course, the key aspect of this work is to consider the purpose evolution analysis as an important tool; but, how can we measure purpose? Since, purpose is something close to the ideas, or underlying motivations of every member, it is not simple to answer this question.

Since, purpose is, from dictionary, "what something is used for". We propose to use goals as a measure of purpose accomplishment. Using this idea, we can measure if a VCoP fulfills a purpose, measuring how well members' contributions accomplish a set of goals previously defined by the owners, managers or experts of the community.

Goals definition must be performed based on interviews or surveys to community experts or administrators ([2,9,10]). Definition of goals consist of a series of phrases. These phrases respond to the question "what the community is for?". This paper only evaluates the goals from community administrators' view point. Although, same process apply for community members' goals analysis.

Afterwards, we need to select a classification or clustering algorithm in order to perform a text mining algorithm to find interesting patterns. It is expected patterns found provide useful information for administrators and experts in order to decide how to enhance the community (add new forums, erase forums, find trends, etc) based on goals fulfillment.

We selected a concept-based text mining since the adaption from concepts to goals is straightforward. This approach will be explained in next section.

As a last consideration, this technique allows us to study the goals' fulfillment through time. Therefore, we can show interesting information of how the VCoP's purpose evolves. Thus, providing useful and objective information to community experts. Without this tool, they only have an intuition on how community has evolved and if the information contained in the community forums is truly accomplishing the purpose of the community.

3.2 Concept-Based Web Text Mining

Concept-based text mining is a data mining approach based on fuzzy sets and fuzzy logic theory.

We based our work in Loh's work presented in [2] and [8]. Loh's proposal is to use a *fuzzy reasoning* model to decide wether a concept is expressed by a web document or not. This way, after the application of the reasoning model, we have classified all documents by its concepts. To do so, we compute the degree of possibility that a concept is related to a web document. In our case, web documents are posts in a VCoP's forum.

This work use goals as a way to evaluate the purpose of VCoP. Therefore, how to introduce goals in the mining process is a key issue. This is why we use the concept based web text mining, this approach allows us to use concepts as goals, then based on the goals definition, the algorithm can classify web documents by the accomplishment of such goals. In the following we will talk about goals instead of concepts.

Fuzzy logic for goals classification. In fuzzy logic, linguistic variables (LV) are not numbers but words or sentences in natural language. These variables are more complex but less precise. Let u be a LV, we can obtain a set of terms $T(u)$ which cover its universe of discourse U. e.g. $T(taste) = \{sweet, salad, acid, bitter, bittersweet\}$

In order to use LV for goals classification, we assume that a community posts can be represented as a fuzzy relation $[Goals \times Posts]$ also called $[G \times P]$. Which is a matrix where each row is a goal and every column is a post in the VCoP. To obtain such matrix we can rewrite this relation in a more convenient manner in Eq.(1) ([2]). In this expression we call "Terms" the words that can be used to define a concept and we write "WP" to refer any word inside a Web Page. In Eq.(1) the symbols "×" and "⊗" represent the fuzzy relation and fuzzy composition respectively.

$$[Goals \times Posts] = [Goals \times Terms] \otimes [Terms \times Posts] \tag{1}$$

As defined above, let P the total amount of posts in a VCoP and W the total number of different words among all of these posts, G the total number of goals defined for the community in study. Then we can characterize the matrix $[G \times P]$ by its membership function shown in Eq.(2), where $\mu_{G\times P} = \mu_{G\times T \otimes T \times P}$ represents the membership function of the fuzzy composition in Eq.(1). The membership values are between 0 and 1.

$$\mu_{G\times P}(x,z) = \begin{pmatrix} \mu_{1,1} & \mu_{1,2} & \cdots & \mu_{1,P} \\ \mu_{2,1} & \mu_{2,2} & \cdots & \mu_{2,P} \\ \vdots & \vdots & \vdots & \vdots \\ \mu_{G,1} & \mu_{G,2} & \cdots & \mu_{G,P} \end{pmatrix} \tag{2}$$

The composition of fuzzy relations is performed using Nakanishi's fuzzy compositional rule Eq.(3). In Eq.(3), let $Q(U,V)$ and $Z(V,W)$ be two fuzzy relations which share a common set V. Let $\mu_Q(x,r)$ with $x \in U \wedge r \in V$ and $\mu_z(r,y)$ with $r \in V \wedge y \in W$ membership functions for Q and Z respectively. Then we can write the compositional rule as shown in Eq.(3). Where $\bigvee$ is the limited Sum $= min(1, x+r)$ and $\wedge$ is the algebraic product $= (x * r)$.

$$\mu_{Q \circ Z} = \bigvee \{\mu_Q(x, r) \wedge \mu_Z(r, y)\} \tag{3}$$

There are several alternatives to perform the fuzzy composition, [3] performed a study between six different reasoning models. One important issue that must be considered is that even if some terms are not present in a post, the degree of that post to express a specific goal should not suffer alterations. This is a reason to use Nakanishi's compositional rule. However, other rules could be used as well.

Identification and Definition of Goals. In order to apply the above proposal, we need to begin identifying the relevant goals for the study. To do so, we make use of community administrator' knowledge whom identify which are the most interesting goals to describe the VCoP's purpose. Subsequently, every goal is represented as a list of terms (assuming that a goal is a LV). We used synonyms, quasi-synonyms, antonyms,etc. using also the administrators.

We realize that several important terms are produced by slang words. For example, the word *transformator* in spanish "transformador" is used as "transformador", "trafo" or "transf" very commonly. Thus, human definition is useful to enhance goals definitions.

Afterwards, we need to define the membership values for the fuzzy relations $[Goals \times Terms]$ and $[Terms \times Posts]$. We used relative frequency of terms in a community post to represent the membership values of matrix $[Terms \times Posts]$.

More difficult is to define $[Goals \times Terms]$ values. We performed this task by asking the community experts to assign these values. For every goal which is the degree that a term has to represent that specific goal. To do so, he compared two terms each time and gave a value between 0 and 1. For example, a synonym can receive a value near 1; a quasi-synonym, may receive a value near between 0.75 and 0.95; an antonym can be set to 0, etc. This method is an indirect method with three experts.

Finally, we obtained the fuzzy relation $\mu_{G \times P}(x, z)$ applying Eq.(3). In Table 1 we present a column of matrix $\mu_{G \times P}(x, z)$, which represents the goals classification for post 4235.html from VCoP. From this Table we can say that post 4325.*html* have a strong relation with the goal 1 and goal 2, almost no relation with goals 3, 4 and 5.

Table 1. List of goals and membership values to represent post "4235.htm"

Goals	$\mu_{G \times P}$
Goal 1	0.88
Goal 2	0.72
Goal 3	0
Goal 4	0.12
Goal 5	0.01
....	

4 A Real Application

We performed the experiments into the web site of plexilandia.cl virtual community. Based on interviews to administrators of this VCoP and a preliminary study of community activity we describe the community used.

4.1 The Community

Plexilandia is a VCoP formed by a group of people who have met towards the building of music effects, amplifiers and audio equipment (like "Do it yourself" style). In the beginning was born as a community for share common experiences in the construction of plexies[1]. Today, plexilandia count more than 2000 members in more than 6 years of existence. All these years they have been shearing and discussing their knowledge about building their own plexies, effects. Besides, there are other related topics such as luthier, professional audio, buy/sell parts.

Although, they have a web page with basic information of community, most of their members' interactions are produced by the discussion forum.

In the beginning the administration task was performed by only one member. Today, this task is performed by several administrators (in 2008 they count with 5 administrators). In fact, the amount of information generated weekly makes impossible to let the administration task in just one admin.

During six years of life, this community has undergone a great sustained growth in members' contributions . The vision of administrators and experts about the community is based mostly by experience and time participating in the community. They also have some basic and global measures. For example, total number of posts, connected members, etc. However, the don't have: members browsing behavior information, members publications' quality and how they contribute to purpose of the community.

4.2 Concept-Based Text Mining Application

First, we selected data from october 2002 to june 2008 and we perform text preprocessing the text to eliminate HTML, Javascript and other programed codes.

In order to apply concept-based text mining approach we need to define concepts or in our case, goals. We defined six different goals with the help of community experts. These where used as input in the concept-based text mining approach. The algorithm took less than five minutes to finish the classification process.

4.3 Analysis of Results

Results obtained where included in a web report. This report allows administrators to understand how the purpose of plexilandia evolves. Report includes

[1] "Plexi" is the nickname given to Marshall amp heads model 1959 that have the clear perspex (a.k.a plexiglass) fascia to the control panel with a gold backing sheet showing through as opposed to the metal plates of the later models.

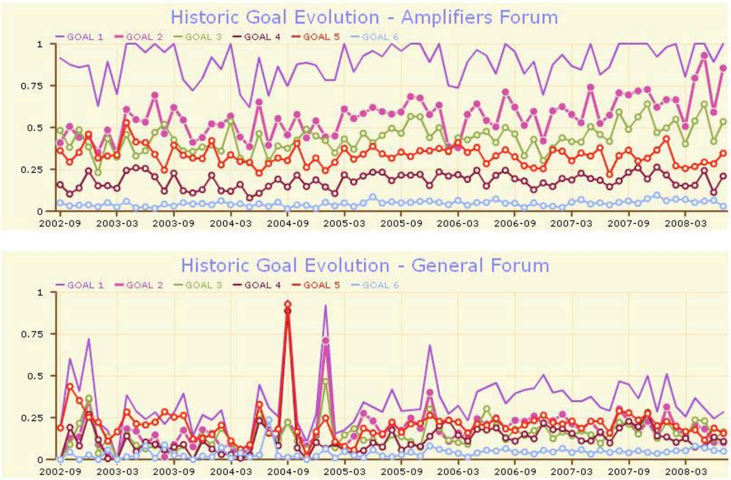


Fig. 1. Goals evolution of two forums

a graph for each forum in the community. This graph represents all goals with a different color. Then, every goal is expressed by its membership value, which means how close is the forum respect of the goals. This can be interpreted as degree of accomplishment of a goal by the forum. If a goal has a value near 1, it indicates that the posts in that forum contributes to the accomplishment of that goal. On the other hand, a value near 0, means that goal doesn't help to accomplish such goal.

For example, in Figure 1 it is possible to observe the forum of "amplifiers" and the level of accomplishment of every defined goal. It is possible to observe that: (i) this form strongly support goal 1, (ii) in last months this forum has experienced a trend to growth in the accomplishment of goal 2, and (iii) the accomplishment of other goals is much lesser. This analysis allows to discover an important conclusion from administration perspective: historically, in "amplifiers" forum the main topic was amplifiers; however, last months there is a trend to talk about related topics, such as music effects (like guitar effects). Therefore, this analysis is an objective tool to show that situation.

Moreover, the report allows to identify certain anomalies, such as, estrange peaks. Administrators, study these peaks and they found perfect sense with particular situations that happened in the forums those months.

4.4 Results' Evaluation

The main objective of these section is to evaluate the validity of experimental results. To do so, we performed interviews to community administrators who have analyzed and validated obtained results.

The importance of this evaluation is based in experts or administrators knowledge and experience about plexies and the community through 6 years. Therefore, they can validate the results but also they can quantify if a result is expected (just by intuition) or they gain an important peace of new knowledge to understand the community.

In addition, it has been applied an usability evaluation of the generated report. This was performed to determine the administrators' satisfaction level at the moment of reading the report.

Both evaluations were applied to 3 from 5 administrators. One of them is the community founder of plexilandia.

Usability Evaluation. To perform this evaluation, community administrators had to answer a survey. We asked the level of satisfaction with the generated report. In the survey, it was measured: ease of use, ease of learn, need of help and report clarity.

Survey results show that report is easy to read and learn but requires of help when reading for first time. This is also related with administrators' previous experience using web reports (only one of them had previous experience).

Since, this is the first report generated and it is oriented only to administrators, we only focused the evaluation in the satisfaction level achieved. In the near future we expect to measure reports' efficiency and efficacy. Then, we pretend to publish results for community members.

Validity of Results. Community administrators had to quantify each result obtained and showed in the report. This is the quantification of anomalies in previous section and a quantification of identified goals evolution.

We used a three points evaluation scale: (1) not expected result; complete surprise, without a clear cause; (2) not expected result, but cause would be known; (3)expected result, known cause.

From 14 anomalies detected, 8 were expected results, and 3 represent situations completely unexpected. From the 30 identifies behavior patterns, 20 represent expected results, and only 2 are surprise situations completely.

5 Conclusion

This work has shown that community administrators are almost blind when enhancing a social network. Moreover, we have also shown that common analysis based just on social network relations are not enough.

We have proved that using the community experts or administrators combined with a data mining approach could provide much more objective and rich information. Which may be used to enhance the virtual community of practice.

Besides, we have proposed the analysis of purpose evolution of a VCoP based on goals definition as the key to the application of data mining analysis into the analysis of VCoP. This way, community experts or administrators count with objective information.

We have successfully used the purpose evolution analysis in a real VCoP with more than 2000 members and 6 years.

We think that although results are promising, it is needed more work in this direction, to find modern tools to help managers, experts or administrators to enhance their communities.

References

1. Koh, J., Kim, Y., Butler, B., Bock, G.: Encouraging participation in virtual communities. Communications of the ACM (January 2007)
2. Loh, S., de Oliveira, J., Gameiro, M.: Knowledge discovery in texts for constructing decision support systems. Applied Intelligence (December 2003)
3. Nakanishi, H., Turksen, I., Sugeno, M.: A review and comparison of six reasoning methods. Fuzzy Sets and Systems (January 1993)
4. Pal, S., Talwar, V., Mitra, P.: Web mining in soft computing framework: relevance, state of the art and future directions. Neural Networks (December 2002)
5. Perkowitz, M., Etzioni, O.: Towards adaptive web sites: Conceptual framework and case study. Artificial Intelligence (December 2000)
6. Preece, J.: Etiquette, empathy and trust in communities of practice: Stepping-stones to social capital. Journal of Universal Computer Science (January 2004)
7. Preece, J., Maloney-Krichmar, D.: Online communities: Focusing on sociability and usability. In: Handbook of Human-Computer Interaction (January 2003)
8. Ríos, S.A.: A study on web mining techniques for off-line enhancements of web sites. Ph.D Thesis, p. 231 (September 2007)
9. Ríos, S.A., Velásquez, J.D.: Semantic web usage mining by a concept-based approach for off-line web site enhancements. In: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence (2008)
10. Ríos, S.A., Velásquez, J.D., Yasuda, H., Aoki, T.: Using a self organizing feature map for extracting representative web pages from a web site. International Journal of Computational Intelligence Research (IJCIR) 2, 159–167 (2006)
11. Shummer, T.: Patterns for building communities in collaborative systems. In: Proceedings of the 9th European Conference on Pattern Languages and Programs (2004)
12. Wellman, B., Salaff, J., Dimitrova, D., Garton, L.: Computer networks as social networks: Collaborative work, telework, and virtual community. Annual Reviews in Sociology (January 1996)
13. Wenger, E.: Communities of practice: Learning, meaning, and identity (January 1999), books.google.com
14. Wenger, E., McDermott, R., Snyder, W.: Cultivating communities of practice: A guide to managing knowledge. Harvard Business School Press (2002)

Discovering Networks for Global Propagation of Influenza A (H3N2) Viruses by Clustering*

Kazuya Sata[1], Kouichi Hirata[1], Kimihito Ito[2], and Tetsuji Kuboyama[3]

[1] Department of Artificial Intelligence, Kyushu Institute of Technology
Kawazu 680-4, Iizuka 820-8502, Japan
{sata,hirata}@dumbo.ai.kyutech.ac.jp

[2] Research Center for Zoonosis Control, Hokkaido University
North 20, West 10 Kita-ku, Sapporo 001-0020, Japan
itok@czc.hokudai.ac.jp

[3] Computer Center, Gakushuin University
Mejiro 1-5-1, Toshima, Tokyo 171-8588, Japan
ori-kes09@tk.cc.gakushuin.ac.jp

Abstract. In this paper, we present a method of discovering networks for modeling global propagation of influenza A (H3N2) viruses using a clustering algorithm. First, we find the clusters for every region by using an agglomerative hierarchical clustering with complete linkage. Next, we collect similar virus clusters over all regions. Finally, by comparing the occurrence year of the similar clusters, we construct a directed graph as a propagation network among these virus clusters.

1 Introduction

Influenza A (H3N2) virus is currently the major cause of human influenza morbidity and mortality worldwide. Therefore, it is an important task to discover the epidemics networks for global propagation of influenza A (H3N2) viruses. Russell *et al.* [5] have studied this task based on antigenic and genetic analysis, and pointed out that, in the network for global propagation, Asia is a starting region, Oceania is an ending region from Asia, and South America is also an ending region from Asia through Europe and North America.

In this paper, we present an alternative method of discovering the networks for the global propagation of influenza A (H3N2) viruses from the viewpoint of data mining and machine learning. The idea of our method is to construct the networks from similar clusters over all regions obtained by clustering [2,3].

In our method, first, we apply an *agglomerative hierarchical clustering with complete linkage* [2,3] to amino acid sequences for every region. As a result, we obtain the set of clusters for every region.

Next, we compute the *characteristic sequence* for every cluster. Then, we collect the *tuple of similar clusters over all regions* such that, for every pair of

* This work is partially supported by Grand-in-Aid for Scientific Research 19300046, 20500126 and 21500145 from the Ministry of Education, Culture, Sports, Science and Technology, Japan.

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 490–497, 2009.

clusters in a tuple, the Hamming distance (*cf.*, [1]) between their characteristic sequences is small.

Finally, from the set of all tuples of similar clusters, we construct a directed graph as a network for global propagation. Every vertex of the graph represents a region. Also, for every tuple q, if clusters c_i and c_j in q from regions i and j $(i \neq j)$, respectively, are similar, and the occurrence year of the characteristic sequence of c_i is smaller than one of c_j, then we add an arc $i \rightarrow j$ from i to j to the set of arcs.

This paper is organized as follows. In Section 2, we prepare some notions necessary for the later discussion, including an agglomerative hierarchical clustering with complete linkage [2,3]. In Section 3, we present our method of discovering propagation networks from the tuples of similar clusters over all regions. In Section 4, we give experimental results applying our method to amino acid sequences of influenza A (H3N2) viruses, provided from NCBI [4], according to 5 regions as Asia, North America, Europe, South America and Oceania.

2 Agglomerative Hierarchical Clustering

In this paper, we use an amino acid sequence of influenza A (H3N2) viruses (a *sequence*, for short) as a sequence with the fixed length $n(= 328)$, and denote it by lower small letters such as x, y, z. The sequence consists of the following 20 amino acids:

$$E = \{\mathtt{A}, \mathtt{R}, \mathtt{N}, \mathtt{D}, \mathtt{C}, \mathtt{Q}, \mathtt{E}, \mathtt{G}, \mathtt{H}, \mathtt{I}, \mathtt{L}, \mathtt{K}, \mathtt{M}, \mathtt{F}, \mathtt{P}, \mathtt{S}, \mathtt{T}, \mathtt{W}, \mathtt{Y}, \mathtt{V}\}.$$

We call such an E the set of *attribute values*. For a sequence x, $x[i]$ $(1 \leq i \leq n)$ denotes the *i-th attribute* of x. Also every x has an *occurrence year* $o(x)$.

Let X be the set of sequences. For $e \in E$ and $1 \leq i \leq n$, let $p_i(e)$ be the probability of the occurrence of e in the i-th attribute of the elements of X. Then, the *entropy* $I_X(i)$ of the i-th attribute of X (*cf.* [2]) is defined as follows.

$$I_X(i) = -\sum_{e \in E} p_i(e) \log_2 p_i(e).$$

The *Hamming distance* $h(x, y)$ between $x \in X$ and $y \in X$ (*cf.*, [1]) is the total number of different attributes, that is, $h(x, y) = \sum_{i=1}^{n} \delta(x[i], y[i])$ for a Kroonecker delta δ. In this paper, we adopt the *Hamming distance with the entropy threshold* σ as the dissimilarity measure to cluster the sequences as follows.

For $x, y \in X$, we define the function $l_i(x, y, \sigma)$ as follows.

$$l_i(x, y, \sigma) = \begin{cases} 1, \text{ if } I_X(i) > \sigma \text{ and } x[i] \neq y[i], \\ 0, \text{ otherwise.} \end{cases}$$

Then, we define the *Hamming distance* $h_\sigma(x, y)$ *under entropy threshold* σ between $x \in X$ and $y \in X$ as $h_\sigma(x, y) = \sum_{i=1}^{n} l_i(x, y, \sigma)$.

As a clustering algorithm, in this paper, we adopt an *agglomerative hierarchical clustering with complete linkage* [2,3] described as Algorithm 1. Here, the complete linkage means that the maximum value is applied to a pairwise dissimilarity measure (in line 12). The input of Algorithm 1 is an object X and a dissimilarity measure d. (In this paper, we set X to the set of sequences and d to h_σ.) On the other hand, the output of Algorithm 1 is a *dendrogram* $D = (V, A)$ (as a directed graph), where V is the set of vertices of the form v_d and d is a dissimilarity measure between two sequences.

```
    procedure AHC(X,d)
        /* $X = \{x_1, \dots, x_m\}$, $d$: dissimilarity measure */
1       $\mathcal{C} \leftarrow \emptyset$; $k \leftarrow m$; $V \leftarrow \emptyset$; $A \leftarrow \emptyset$;
        /* Initialization */
2       for $i = 1$ to $m$ do
3           $c_i \leftarrow \{x_i\}$; $\mathcal{C} \leftarrow \mathcal{C} \cup \{c_i\}$; $V \leftarrow V \cup \{v_0^i\}$;
4           for $j = i + 1$ to $m$ do
5               $d(c_i, c_j) \leftarrow d(x_i, x_j)$;
6       while $k \neq 1$ do
7           $d(c_q, c_r) \leftarrow \min\{d(c_i, c_j) \mid c_i, c_j \in \mathcal{C}, i \neq j\}$; $d \leftarrow d(c_q, c_r)$;
8           $V \leftarrow V \cup \{v_d\}$; $A \leftarrow A \cup \{(v_q, v_d), (v_r, v_d)\}$;
9           $c' \leftarrow c_p \cup c_r$;
10          $\mathcal{C} \leftarrow (\mathcal{C} - \{c_q, c_r\}) \cup \{c'\}$;
11          foreach $c \in \mathcal{C} - \{c'\}$ do
12              $d(c, c') = \max\{d(x, y) \mid x \in c,\ y \in c'\}$;
13          $k \leftarrow k - 1$;
14      return a dendrogram $D = (V, A)$;
```

Algorithm 1. *AHC* [2,3].

From a dendrogram $D = (V, A)$ given by Algorithm 1, we can obtain the set of clusters under the *dissimilarity threshold* t as follows. Let D_t be the restriction of D under t, that is, $V_t = \{v_d \in V \mid d \leq t\}$, $A_t = \{(v_d, v_e) \in A \mid d \leq e \leq t\}$ and $D_t = (V_t, A_t)$. Also let $S \subseteq V_t$ be the set of all sinks (that is, vertices with outdegree 0) in D_t, and, for every $s \in S$, $V_0(s)$ the set of all sources (that is, vertices with indegree 0, which is of the forms v_0^i) with the sink s. Then, we can construct a cluster as a set $\{x_i \in X \mid v_0^i \in V_0(s)\}$ for every sink $s \in S$ of D_t.

As stated above, our method is necessary to use two parameters of the entropy threshold σ and the dissimilarity threshold t. In the following, we also introduce another two parameters, the *cluster threshold* ω and the *comparison threshold* s, to compare the clusters.

Let C be a set of clusters and c a cluster in C. Also let $c[i]$ be the most frequent attribute value in $x[i]$ for every $x \in c$. Then, we formulate a *characteristic sequence* $cs(c)$ of a cluster c *under the cluster threshold* ω such that the i-th attribute $(cs(c))[i]$ of $cs(c)$ is defined by:

$$(cs(c))[i] = \begin{cases} c[i], \text{ if } I_c(i) \leq \omega, \\ *, \quad \text{otherwise.} \end{cases}$$

Here, we assume that $cs(c)$ has an *occurrence year* $o(cs(c))$ as the minimum occurrence year $o(x)$ of an element x in the cluster c.

Let C_1 and C_2 be two sets of clusters. Then, we say that two clusters $c_1 \in C_1$ and $c_2 \in C_2$ are *similar under the comparison threshold* s if $h(cs(c_1), cs(c_2)) \leq s$.

3 Discovering Networks for Global Propagation

In this section, we present our method to discover the networks for the global propagation of influenza A (H3N2) viruses based on an agglomerative hierarchical clustering with complete linkage stated in Section 2. First, we describe the method of collecting the *tuples of similar clusters over all regions* as follows.

1. Set four parameters of the entropy threshold σ, the dissimilarity threshold t, the cluster threshold ω, and the comparison threshold t. Also let X_i be the set of sequences from a region i for $1 \leq i \leq r$.
2. For every i ($1 \leq i \leq r$), run the algorithm $AHC(X_i, h_\sigma)$ by using h_σ. Let C_i be the set of clusters obtained from $AHC(X_i, h_\sigma)$ under t.
3. For every i ($1 \leq i \leq r$) and $c \in C_i$, compute a characteristic sequence $cs(c)$ of c under ω.
4. Let $Q = \emptyset$. For every $q = (c_1, \ldots, c_r) \in C_1 \times \cdots \times C_r$, check whether or not $h(cs(c_i), cs(c_j)) \leq s$ for every i and j ($1 \leq i < j \leq r$). If so, then add q to Q. We call such a q a *tuple of similar clusters over all regions* $X_1 \cup \cdots \cup X_r$ *under* s.

Example 1. Assume that the number of regions is 3, and consider the three sets X_1, X_2 and X_3 shown in Table 1, where the first and the second columns are an id and an occurrence year, respectively. Also we set four parameters to $t = 1$, $\sigma = 0.1$, $\omega = 0.1$, $s = 1$.

From the result of the algorithm AHC under $\sigma = 0.1$ and $t = 1$, we obtain the following sets C_1, C_2 and C_3 of clusters of X_1, X_2 and X_3, respectively.

$$C_1 = \{\{a, b\}, \{c\}, \{d, e\}\}, C_2 = \{\{f\}, \{g, h\}, \{i, j\}\}, C_3 = \{\{k\}, \{l, m\}, \{n, o\}\}.$$

From C_1, C_2 and C_3, we can obtain the characteristic sequence of every cluster under $\omega = 0.1$ shown in Table 2.

Finally, we obtain two tuples of similar clusters over $X_1 \cup X_2 \cup X_3$ under $s = 1$ as $(\{a, b\}, \{g, h\}, \{n, o\})$ and $(\{c\}, \{f\}, \{l, m\})$, the characteristic sequences and occurrence years of which are shown in Table 3.

Table 1. The sets X_1, X_2 and X_3 of sequences in Example 1

X_1			X_2			X_3		
id	year	sequence	id	year	sequence	id	year	sequence
a	2002	Q L P A S N T Q K S	f	2002	Q L N T S N T P K S	k	2003	Q L P A S D T Q K S
b	2002	Q L S A S N T Q K S	g	2003	Q L S A S N T Q K S	l	2003	Q L N T S N T Q K S
c	2002	Q L P T S N T Q K S	h	2003	Q L Q A S N T Q K S	m	2004	Q L Q T S N T Q K S
d	2004	Q A S A N N T Q K S	i	2004	Q L Q A N T T Q K S	n	2005	Q L S A S N T Q K S
e	2005	Q A S A N T T Q K S	j	2005	Q L Q A N S T Q K S	o	2005	Q L S A S N T Q K S

Table 2. The characteristic sequences for C_1, C_2 and C_3

C_1		
$\{a, b\}$	2002	QL∗ASNTQKS
$\{c\}$	2002	QLPTSNTQKS
$\{d, e\}$	2004	QASAN∗TQKS

C_2		
$\{f\}$	2002	QLNTSNTPKS
$\{g, h\}$	2003	QL∗ASNTQKS
$\{i, j\}$	2004	QLQAN∗TQKS

C_3		
$\{k\}$	2003	QLPASDTQKS
$\{l, m\}$	2003	QL∗TSNTQKS
$\{n, o\}$	2005	QLSASNTQKS

Table 3. The tuples of similar clusters over all regions in Example 1

C_1:	$\{a, b\}$	2002	QL∗ASNTQKS
C_2:	$\{g, h\}$	2003	QL∗ASNTQKS
C_3:	$\{n, o\}$	2005	QLSASNTQKS

C_1:	$\{c\}$	2002	QLPTSNTQKS
C_2:	$\{f\}$	2002	QLNTSNTQKS
C_3:	$\{l, m\}$	2003	QL∗TSNTQKS

Next, from the set Q of tuples of similar clusters over all regions obtained by the above method, we construct a directed graph $G = (V, A)$ as the network for global propagation of influenza A (H3N2) viruses as follows: V is the set of regions, and A consists of an arc $i \rightarrow j$ from i to j if $o(cs(c_i)) < o(cs(c_j))$ for every tuple $q = (c_1, \dots, c_r) \in Q$ and two regions i and j. Furthermore, let $w_{i \rightarrow j}$ be the number of elements in Q such that $o(cs(c_i)) < o(cs(c_j))$ for $c_i, c_j \in q$ and $q \in Q$. Then, the *weight* of an arc $i \rightarrow j$ from i to j is $w_{i \rightarrow j}/|Q|$. Note that $w_{i \rightarrow j} + w_{j \rightarrow i} \leq |Q|$ but $w_{i \rightarrow j} + w_{j \rightarrow i}$ is not always equal to $|Q|$, because of the existence of the case that $o(cs(c_i)) = o(cs(c_j))$.

Example 2. Consider the tuple $(\{a, b\}, \{g, h\}, \{n, o\})$ of similar clusters over $X_1 \cup X_2 \cup X_3$ in Example 1. Since $o(cs(\{a, b\})) = 2002$, $o(cs(\{g, h\})) = 2003$ and $o(cs(\{n, i\})) = 2005$ in Table 3, we can obtain the arcs $1 \rightarrow 2$ and $2 \rightarrow 3$ such that the set of regions is $\{1, 2, 3\}$.

On the other hand, consider the tuple $(\{c\}, \{f\}, \{l, m\})$ of similar clusters over $X_1 \cup X_2 \cup X_3$ in Example 1. Since $o(cs(\{c\})) = 2002$, $o(cs(\{f\})) = 2002$ and $o(cs(\{l, m\})) = 2003$ in Table 3, we can obtain the arcs $1 \rightarrow 3$ and $2 \rightarrow 3$ such that the set of regions is $\{1, 2, 3\}$.

Hence, we can obtain arcs of $1 \rightarrow 2$ with weight $1/2$, $1 \rightarrow 3$ with weight $1/2$, and $2 \rightarrow 3$ with weight $2/2$.

4 Experimental Results

In this section, we give experimental results of networks for the global propagation of influenza A (H3N2) viruses using our method in Section 3. We adopt the amino acid sequences of influenza A (H3N2) viruses in years between 2002 and 2006 provided from NCBI [4]. The number of sequences is 1825. Table 4 describes the global distribution of such sequences.

Table 4. The global distribution of amino acid sequences of influenza A (H3N2) viruses

year	Asia	North America	Europe	South America	Oceania
2002	58	90	75	11	111
2003	21	109	207	2	112
2004	58	87	104	14	158
2005	58	52	110	3	122
2006	94	69	55	25	13
total	289	407	551	55	516

Table 5. The number $|Q|$ of tuples of similar clusters over all regions by changing four parameters t, σ, ω and s

t	σ	ω	s	$\|Q\|$
5	0.1	0.6	7	12
6	0.1	0.6	7	3
7	0.1	0.6	7	0
5	0.1	0.6	8	30
6	0.1	0.6	8	16
7	0.1	0.6	8	8
5	0.1	0.6	9	49
6	0.1	0.6	9	44
7	0.1	0.6	9	30

t	σ	ω	s	$\|Q\|$
5	**0.0**	**0.6**	**7**	**9**
5	0.1	0.6	7	12
5	0.2	0.6	7	8
6	0.0	0.6	7	0
6	0.1	0.6	7	3
6	0.2	0.6	7	1
7	0.0	0.6	7	0
7	0.1	0.6	7	0
7	0.2	0.6	7	0

t	σ	ω	s	$\|Q\|$
5	0.1	0.5	7	0
5	**0.1**	**0.6**	**7**	**12**
5	0.1	0.7	7	35
6	0.1	0.5	7	0
6	0.1	0.6	7	3
6	0.1	0.7	7	21
7	0.1	0.5	7	0
7	0.1	0.6	7	0
7	0.1	0.7	7	11

t	σ	ω	s	$\|Q\|$
5	0.1	0.6	7	12
5	0.1	0.6	8	30
5	0.1	0.6	9	49
6	**0.1**	**0.6**	**7**	**3**
6	0.1	0.6	8	16
6	0.1	0.6	9	44
7	0.1	0.6	7	0
7	0.1	0.6	8	8
7	0.1	0.6	9	30

When we apply our method to the above data, we set the number n of regions to 5, and X_1, X_2, X_3, X_4 and X_5 to the sequences from Asia, North America, Europe, South America and Oceania, respectively.

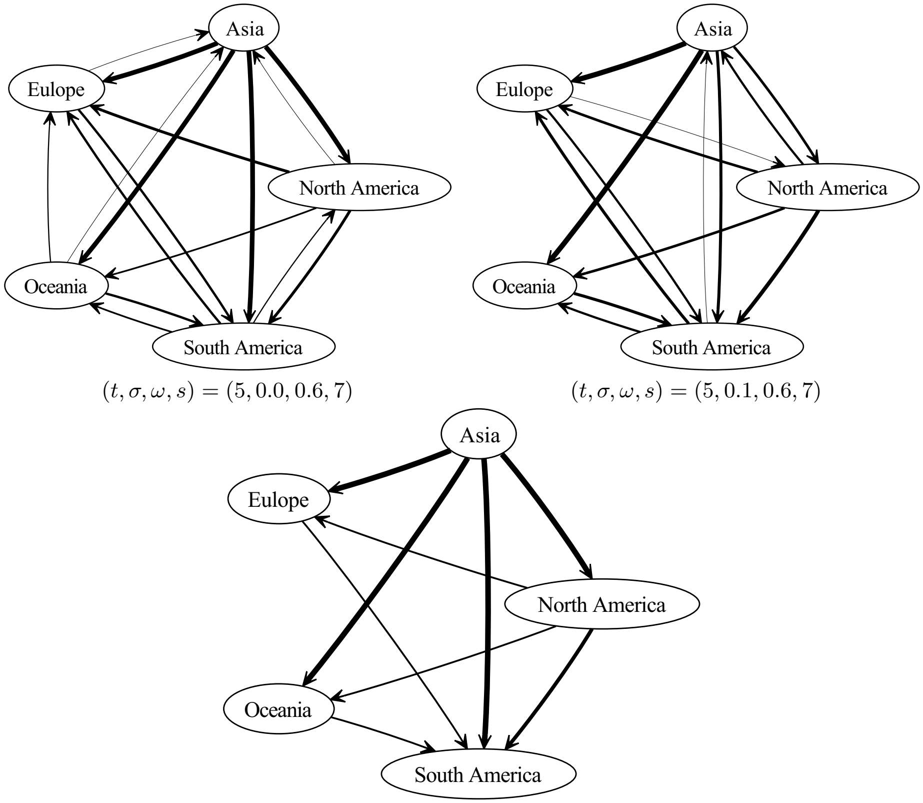
First, we discuss how the number $|Q|$ of tuples of similar clusters over all regions changes when changing four parameters σ, t, ω and t. Table 5 shows that (1) $|Q|$ does not change when changing σ, (2) $|Q|$ increases if ω or s increases, and (3) $|Q|$ decreases if t increases.

The three networks for global propagation are shown in Figure 1 with the parameters (t, σ, ω, s) as $(5, 0.0, 0.6, 7)$ (upper left), $(5, 0.1, 0.6, 7)$ (upper right) and $(6, 0.1, 0.6, 7)$ (lower), where the arc thickness is proportional to the arc weight.

For the case that $(t, \sigma, \omega, s) = (5, 0.0, 0.6, 7)$ in Figure 1 (upper left), Q consists of 9 tuples, and we observe that the influenza A (H3N2) viruses start from Asia and reach at South America through North America. Also we observe the bidirectional propagations between South America and either Europe or Oceania, and no propagation between Europe and Oceania.

For the case that $(t, \sigma, \omega, s) = (5, 0.1, 0.6, 7)$ in Figure 1 (upper right), Q consists of 12 tuples, and we observe the similar propagations of Figure 1 (upper left), except one from North America to Asia.

For the case that $(t, \sigma, \omega, s) = (6, 0.1, 0.6, 7)$ in Figure 1 (lower), Q consists of three tuples, and we observe that the influenza A (H3N2) viruses start at Asia



$(t, \sigma, \omega, s) = (5, 0.0, 0.6, 7)$

$(t, \sigma, \omega, s) = (5, 0.1, 0.6, 7)$

$(t, \sigma, \omega, s) = (6, 0.1, 0.6, 7)$

Fig. 1. The networks for global propagation for $(t, \sigma, \omega, s) = (5, 0.0, 0.6, 7)$ (upper left), $(5, 0.1, 0.6, 7)$ (upper right) and $(6, 0.1, 0.6, 7)$ (lower)

and end at South America through North America, Europe and Oceania, and observe no propagation between Europe and Oceania.

5 Conclusion

In this paper, we have designed the method to discover the networks for the global propagation of influenza A (H3N2) viruses by using a clustering algorithm. Then, as experimental results, we have discovered the networks for propagation starting from Asia and transiting North America, and no propagation between Europe and Oceania, which is the same as Russell's work [5]. On the other hand, our networks contain the bi-directional propagations between South America and either Europe or Oceania, which is different from Russell's work [5].

It is one of the important future works to analyze the reasons why the above difference occurs in more detail. In particular, it is necessary to evaluate covered sequences by a characteristic sequence. Concerned with this work, it is also a future work to evaluate the networks obtained by our method from the antigenic and genetic viewpoints.

Also, it is a future work to improve our method in order to discover the networks with high accuracy, for example, how to select a characteristic sequence and an arc in a network. Furthermore, it is a future work to adopt another clustering algorithm and then to discover the networks, instead of an agglomerative hierarchical clustering with complete linkage.

References

1. Crochemore, M., Hancart, C., Lecroq, T.: Algorithms on strings. Cambridge University Press, Cambridge (2007)
2. Hastie, T., Tibshirani, R., Friedman, J.: The elements of statistical learning. Springer, Heidelberg (2001)
3. Jain, A.K., Murty, M.N., Flynn, P.J.: Data clustering: A review. ACM Computing Survey 31, 264–323 (2000)
4. National Center for Biotechnology Information (NCBI): Influenza virus resources (2008), http://www.ncbi.nlm.nih.gov/genomes/FLU/FLU.html
5. Russell, C.A., Jones, T.C., Barr, I.G., Cox, N.J., Garten, R.J., et al.: The global circulation of seasonal influenza A (H3N2) viruses. Science 320, 340–346 (2008)

Machine Vision Application to Automatic Intruder Detection Using CCTV

Hernando Fernandez-Canque[1], Sorin Hintea[2], John Freer[1], and Ali Ahmadinia[1]

[1] Glasgow Caledonian University, School of Engineering & Computing
Cowcaddens Road, Glasgow. G4 0BA, United Kingdom
hfe@gcal.ac.uk
[2] Technical University of Cluj Napoca, Str. Baritiu, Nr. 26-28, 3400 Cluj Napoca, Romania,
Tel.: +4 064 196 285, Fax: +4 064 191 340

Abstract. The work presented in this paper addresses the application of new technologies to the task of intruder monitoring. It presents an innovative Machine Vision application to detect and track a person in a Closed Circuit Television System (CCTV) identifying suspicious activity. Neural Network techniques are applied to identify suspicious activities from the trajectory path, speed, direction and risk areas for a person in a scene, as well as human posture. Results correlate well with operator determining suspicious activity. The automated system presented assists an operator to increase reliability and to monitor large numbers of surveillance cameras.

Keywords: Machine vision, image processing, CCTV, human posture recognition.

1 Introduction

In recent years, the use of surveillance cameras has increased in popularity. This is partially due to reduction in cost and technological advances. CCTV (Closed Circuit Television) systems have become very popular in observing public places. Current technology makes provision for an operator to examine live surveillance footage from remote locations as they can be transmitted over the internet, cables or wireless mediums. In this paper we want to enhance early work in this field [1] to increase reliability and potential to detect suspicious activity by the studying of human posture and observing full trajectories of people. In this study, work has been carried out with the aim of achieving fully automatic detection of intruders using a static camera and in real time. CCTV has the advantage that relatively large areas can be monitored and intruders can be seen as compared to other detection methods. The main use of CCTV is based on reaction to a past incident by revising image recorded; the aim of our work is to make the use of CCTV more efficient by assessing suspicious activity in an active manner and alert operators to an intrusion. By achieving an automatic detection some problems associated with this type of surveillance can be avoided. It is known that the span of concentration of any operator is very short [2], and there is unreliability due to operator's fatigue and poor detection due to large number of irrelevant images known as Eriksen effect[3]. Our approach produces realistic results with less

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 498–505, 2009.

use of computer power and can cover large analysis of images in one system. Other works carried out in this field use more than one camera [4], 3D images [5] and skeletonization [6] increasing the amount of computer power required. In our work, artificial neural network, ANN, is used as a decision tool based on processed images. The Machine Vision System presented here has been designed to improve the current analysis procedure in order to simplify and improve detection reliability and warning signals.

2 Problems Associated with Automatic Detection

2.1 Auto-focussing

Auto focusing is achieved by computing a focus score, and then trying to maximise this score by adjusting the camera focus control. The focus score provides a measure of the overall focus quality; therefore, in the case of an uneven focus plane this algorithm would find the best compromise. [7]

2.2 Illumination

A very important aspect for a Machine Vision System is the image representation. Colour images can be represented in several different formats. The most common representation of colours is the Red, Green, Blue (RGB) colour space, as most image sensors provide data according to this model. RGB colour plane is suitable for image capture and reproduction, but for feature extraction proved to be inconvenient in this type of application, as histograms are greatly affected by changes in illumination. Problems are noticed when minor changes in the lighting conditions occur. One solution to overcome this problem is the use of a linear transformation from the RGB colour space into an alternative model Hue, Saturation and Lightness (HSL). The advantages over the RGB colour space is illustrated in figure 1 where the histogram for the same frame with 10% change in lighting condition is presented.

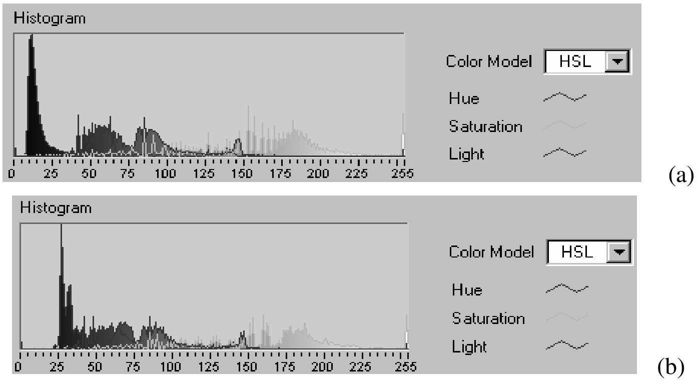

Color Model: HSL

Fig. 1. HSL histogram indicating the number of occurrences of a pixel versus pixel value for (a) test image (b) Test image with 10% change of lighting conditions

From these figures, we can see that the saturation plane is slightly affected and the hue plane is relatively unaffected. HSL will provide a threshold to separate objects of interest reducing the effect of illumination changes.

3 Detection Algorithm

The algorithms presented here detect the presence of a human within CCTV images by comparing images to a pre storage image without intruders present. An alarm can be activated if the object is approaching or is in a risk area; if the subject stops or slows down the algorithm presented can analyse posture for suspicious activity.

3.1 Object Detection and Tracking

This algorithm applies a median filtering to capture image, subtraction from original scene, histogram to select threshold, threshold and binarisation, pixel count to detect subject. A large number of pixel difference will indicate the presence of an intruder. Given the small size of objects in the external scene, additional processing steps are carried out. These steps include analysis of detected clusters yielding data on object size and shape characteristics.

The *shape factor*, F_C, is defined as:-

$$F_C = \frac{L(X)^2}{4\pi . A(X)} \tag{1}$$

where A(X) is the area of the object X and L(X) is the perimeter of the object. It is invariant to rotation, reflection and scaling. It has no dimension and is equal to 1 for a disk. It measures the elongation of an object.

The inertia moments define some global characteristics of the object but it is the first order moments of inertia that define the *barycentre*. They are defined in the discrete case as:-

$$M_{1x} = \frac{1}{A(X)}\sum_{X} x_i \qquad M_{1y} = \frac{1}{A(X)}\sum_{X} y_j \tag{2}$$

where M_{1x} is the first moment of inertia in the *x* plane. M_{1y} is the first moment of inertia in the *y* plane. (x_i,y_j) is a point in the object.

The additional series of processing steps for the external scene is shown in figure 2.

The *barycentre* and position of each detected object is passed to the tracking algorithm after every frame acquisition. Subsequent frame acquisitions provide a new positional locator for each detected object. The tracking algorithm computes the linear distance from every initially detected object to every object detected in the subsequent frame acquisition. The shortest distance between each initially detected object and subsequently detected objects is selected. The object that lies the shortest distance from the initial object is then determined to be the same object as in the previous frame. The process is repeated for each frame acquisition thus allowing objects to be tracked. The *Si factor* provides one method for determining that tracked objects between successive frame captures are the same object within the images. The Si factor can be calculated as follows:-

$$Si = \frac{\frac{|A(X_{nI1}) - A(X_{nI2})|}{A(X_{nI1})} \times 100 + \frac{|F_c(X_{nI1}) - F_c(X_{nI2})|}{F_c(X_{nI1})} \times 100}{2} \tag{3}$$

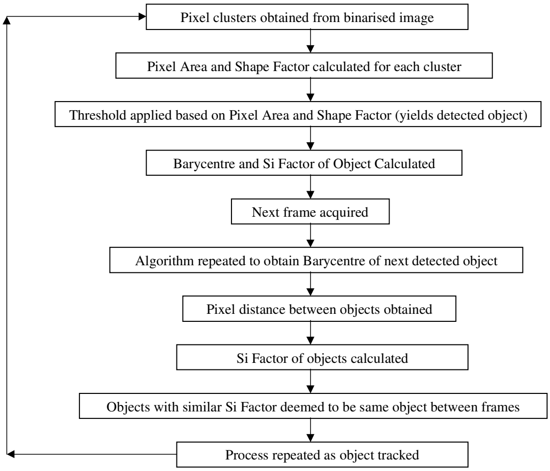

Pixel clusters obtained from binarised image
Pixel Area and Shape Factor calculated for each cluster
Threshold applied based on Pixel Area and Shape Factor (yields detected object)
Barycentre and Si Factor of Object Calculated
Next frame acquired
Algorithm repeated to obtain Barycentre of next detected object
Pixel distance between objects obtained
Si Factor of objects calculated
Objects with similar Si Factor deemed to be same object between frames
Process repeated as object tracked

Fig. 2. Processing steps external scene

where $A(X_{nI1})$ is the *area* of object X_n in Image 1. $A(X_{nI2})$ is the *area* of object X_n in Image 2. $F_c(X_{nI1})$ is the *shape factor* of object X_n in Image 1. $F_c(X_{nI2})$ is the *shape factor* of object X_n in Image 2

The *Si factor* is calculated for all objects detected and provides a confidence measure to determine that objects tracked between images are the same object. Objects detected between successive frames which have a Si factor which lies above the threshold can be assumed to be different objects. This provides the capability for the tracking algorithm to detect when an object has been lost rather than tracking the incorrect object.

3.2 Detection and Suspicious Activity

The identification of position and dimensions of a bounding box containing the object on interest is obtained to allow investigation of suspicious activity. Let x_2 be the x co-ordinate of the non-zero pixel with the lowest x co-ordinate value in the binary image. Let x_2 be the x co-ordinate of the non-zero pixel with the highest x co-ordinate value in the binary image. Let y_2 be the y co-ordinate of the non-zero pixel with the lowest y co-ordinate value in the binary image. Let y_2 be the y co-ordinate of the non-zero pixel with the highest y co-ordinate value in the binary image. The co-ordinates of the four corners of the object-bounding box are therefore:- (x_1,y_1), (x_2,y_1), (x_1,y_1), (x_2,y_2).

The detection of suspicious activity will require further processing on the image containing the object of interest within the bounding box. A novel process to eliminate noise helps the automatic selection of thresholds of the object of interest using the following algorithm: (i) a buffer copy of the image to be cleaned up is generated. (ii) two successive erode functions are applied on the original image. (iii) all pixels from the buffer copy 8-connected to the non-zero pixels from the image are added to the image. (iv) step (iii) is repeated until no pixel is added. The next step is to reinsert the missing pixels within the object boundary. A closing algorithm is performed, using a kernel size of 5. A NOT function is performed. The result is an image which has a value of 0 associated with all objects, a value of 1 associated with the background and a value greater than 1 for every hole in the objects. By replacing the values greater than 1 with 0 and negating the image again, we achieve filling of holes. All objects too small are eliminated. This is achieved using the same algorithm as for binary noise removal, but with 7 erosion functions applied.

After binarisation and elimination of noise, the object of interest is ready to be analysed suspicious activity using ANN. The ANN implemented is a Multi Layer Perceptron with one hidden layer.

4 Results

Knowledge of the scene is an important factor in determining a risk factor of the different areas. A sequence of images is captured from an external scene. Figure 3 shows a test scene used. This scene has four distinct areas: Area 1 pathways, Area 2 car park, Area 3 exit/entrances, Area 4 perimeter area. Risk Index takes values between 0 and 1, where 0 represent the lowest risk. For this exercise the areas were given the following risk factors: Area 1 risk 0.2, area 2 risk 0.6, Area 3 risk 0.7, and area 4 risk 0.8. The risk factor, the speed and direction of the object of interest will provide information to be inputted to the neural network to decide in a warning.

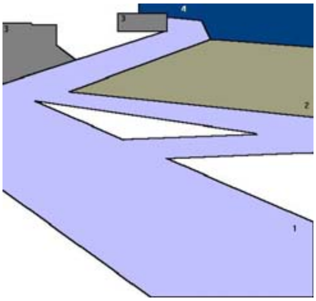


Fig. 3. Scene used with the four areas

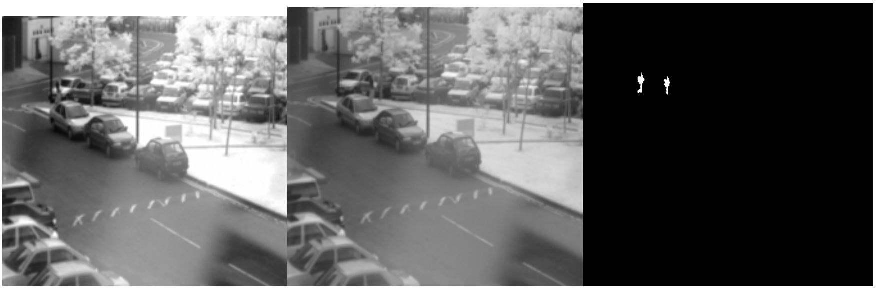
4.1 Tracking

Figure 4 (a) and (b) shows 2 consecutive scenes containing one person, taken with the same camera set up. Each scene is processed under the detection algorithm. After threshold and binarisation and comparing with the original scene a number of clusters are found. The shape and area are calculated for each cluster allowing the elimination of clusters that are not within the shape and area threshold for our object of interest: a person.

Table 1 shows area factor, shape factor, Si factor and position for the cluster detected as a person using the threshold for a person for these two scenes. Other cluster will have large differences in shape, Si factor and area in two consecutives scenes and they will be outside thresholds.

Table 1. Parameters for scene (a) and (b) in fig. 4

	Scene (a)	Scene (b)	Min Threshold
position	102,137	147,139	
Area factor	211	167	150
Shape factor	3.11	2.70	2.5
Si factor	17	17	



(a) (b) (c)

Fig. 4. Two consecutive scenes (a) and (b). In (c) the object of interest is shown for the 2 scenes after detection algorithms have been applied.

Cluster will appear for object of no interest such birds, movements of branches, etc. After the detection and elimination of unwanted noise figure 4(c) shows the detected person of the two scenes superimposed.

4.2 Motion Analysis

Detected objects that have met the size, shape and Si factor criteria are tracked in terms of their position within camera images. The Barycentre of each detected object is passed to the tracking algorithm after every frame acquisition. Subsequent frame acquisitions provide a new position for each detected object. The tracking algorithm computes the linear distance from every initially detected object to every object detected in the subsequent frame acquisition. Figure 5 shows the movement of a person within the segmented areas of the scene analysed.

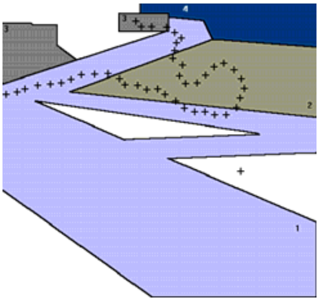


Fig. 5. Human walker within a segmented image

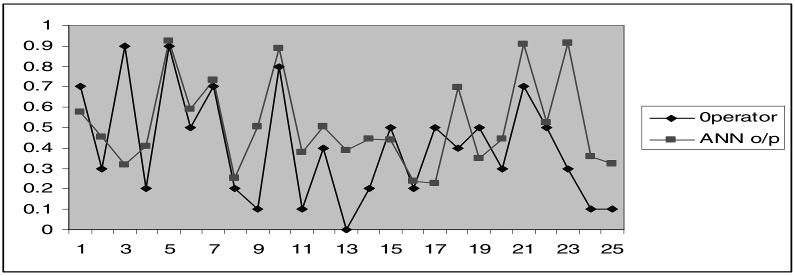


Fig. 6. Response from the ANN and operator to 25 samples

From this tracking scene information of speed, direction and presence in any of the 4 areas was extracted to be presented to a trained ANN for determination of suspicious activity. Figure 6 shows the response of the ANN to 25 path samples of people walking patterns in the different areas of the scene analysed indicating the degree of suspicious activity for each pattern. In the same figure the degree of suspicious activity assessed by human operators is also indicated.

The ANN results of suspicious activity show a good correlation with the human operator response within scenes. An overall error of 10.7% difference between the ANN results and human operator results were found for these experimental results. This provides a good forewarning role allowing further investigation by a human operator.

4.3 Posture Analysis

If the person is not moving in areas that are considered a risk, his or her posture may indicate suspicious activity. This can be assessed from a scene following a bounding box determination and after the object is clean of noise with a well binarised image. Each image is subjected to a reduction algorithm, producing a quantised image, followed by a 16 by 16 data array to be presented to the ANN. The ANN was trained to provide an output equal between 0.88 and 0.95 for a crouching posture. Figure 7 shows results from a trained neural network to a set of 25 images containing humans in crouching position.

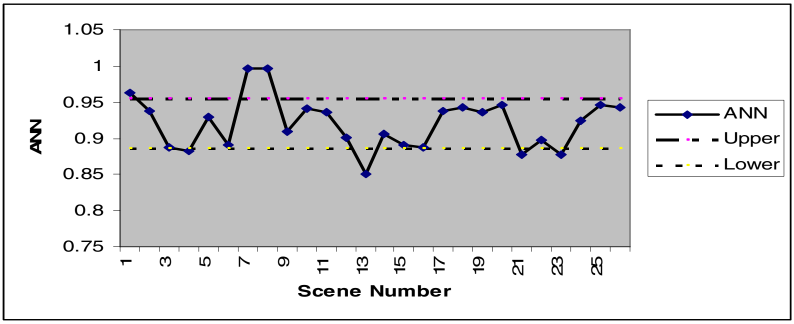


Fig. 7. Neural Network response for crouching postures

The results are close to the expected response, few responses fall outside the range by a small margin with a maximum error of 5%. This can also be used as a warning role.

5 Conclusions

The Machine Vision approach presented in this paper can improve the performance of automated analysis to determine whether suspicious activity is present in an open scene under CCTV surveillance. Working with HSL space reduces the problems associated with change in illumination conditions. The use of the barycentre allows the tracking of an object of interest reducing the amount of data to be processed. Our system can determine the presence and movements of a person and provide indications of suspicious activity based on pathway taken, speed, direction; the system can also provide indication of suspicious activity based on human posture analysis. These enable efficient monitoring and accurate review of scenes by an operator. The system proposed can provide a warning role to reduce the problem of human operator's fatigue and shortened attention span. This greatly increases the ability to carry out properly the task of constant and distant monitoring. The early groundwork proposed show promising results in detection of intruders in a large number of cameras.

References

1. Freer, J.A., Beggs, B.,, Fernandez-Canque, H.L., Chevrier, F., Goryashko, A.: Automatic Recognition of Suspicious Activity for Camera Based Security Systems. In: IEE proceeding of European Convention on Security and Detection, Brighton, vol. (408), pp. 54–58 (1995)
2. Saarinen, J., Julesz, B.: The Speed of Attentional Shifts in the Visual-field. Proceeding Academy of Sciences of the United States of America 88(5), 1812–1814 (1991)
3. Eriksen, C.W., Hoffman, J.E.: Temporal and spatial characteristics of selective encoding from visual displays. Perception & Psychophysics 12, 201–204 (1972)
4. Cucchiara, R., Prati, A., Vezzani, R.: posture classification in a multi-camera environment. In: Proceeding of International Conference on Image Processing, pp. 725–728 (2005)
5. Agarwal, A., Triggs, B.: Recovering 3D Human Pose from Monocular Images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44–58 (2006)
6. Kellokumpu, V., Pietikainen, M., Heikkila, J.: Human Activity Recognition Using Sequence of Postures. In: IAPR Conference on Machine Vision Application, pp. 570–573 (2005)
7. Fernandez-Canque, H., Hintea, S., Csipkes, G., Pellow, A., Smith, H.: Machine Vision Application to the Detection of Micro-organism in Drinking Water. In: Lovrek, I., Howlett, R.J., Jain, L.C. (eds.) KES 2008, Part III. LNCS, vol. 5179, pp. 302–309. Springer, Heidelberg (2008)

A Genetic Algorithm-Based Multiobjective Optimization for Analog Circuit Design

Gabriel Oltean, Sorin Hintea, and Emilia Sipos

Technical University of Cluj-Napoca, Bases of Electronics Departament,
C-tin Daicoviciu 15, 400020 Cluj-Napoca, Romania
{Gabriel.Oltean,Sorin.Hintea,Emilia.Sipos}@bel.utcluj.ro

Abstract. Multiple, often conflicting objectives are specific to analog design. This paper presents a multiobjective optimization algorithm based on GA for design optimization of analog circuits. The fitness of each individual in the population is determined using a multiobjective ranking method. The algorithm found a set of feasible solutions on the Pareto front. Thus, the circuit designers can explore more possible solutions, choosing the final one according to further preferences/constraints. The proposed algorithm was shown to produce good solutions, in an efficient manner, for the design optimization of a CMOS amplifier, for two different sets of requirements.

Keywords: genetic algorithm, multiobjective optimization, Pareto ranking, Pareto front, analog circuit design.

1 Introduction

Designing the analog part of a mixed-signal complex electronic circuit requires a long time of the overall design time, even if the analog part represent only a small fraction of the circuit. Unlike its digital counterpart, the analog design domain is not blessed with powerful tools that simplify the design process [1]. The role of CAD techniques for circuit analysis and optimization became essential to obtain solutions that satisfy the requested performance with the minimum time effort [2]. Due to the complexity of analog circuits, global and local optimization algorithms have to be extensively employed to find a set of feasible solutions that satisfies all the objectives and constraints required by a given application.

Traditional single objective optimization does not allow multiple competing objectives to be accounted for explicitly; moreover they do not give circuit designer the freedom to choose among different, equally feasible solutions. A big step forward in this direction can be achieved using a multiobjective approach. This technique allows different objectives to be treated separately and simultaneously during the optimization process [2].

A multiobjective optimization algorithm should provide a set of nondominated individuals (Pareto front), or optimal solution set. Generating the Pareto set can be computationally expensive and it is often infeasible, because the underlying application prevents exact methods from being applicable. A number of stochastic search strategies such as evolutionary algorithms, tabu search, simulated annealing, and ant

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 506–514, 2009.

colony optimization have been developed [3]. As evolutionary algorithms are assumed to yield good results on complex problems without explicit knowledge of the detailed interdependencies involved, they seem to be a tempting choice [1].

Genetic algorithms (GA) performing multiobjective optimization (MOO) have previously been used in analog circuit design to generate a set of Pareto optimal solutions. In [2] the problem of analog IC design is formulated as a constrained MOO problem defined in a mixed integer/discrete/continuous domain. In [4] analog circuit satisfying a specific frequency response, using free circuit structures and including some parasitic effects, are produced in a single design stage. A coding scheme where the structure of the circuit and parameter values are encoded in a single chromosome, and a multiobjective GA is used in [5] to search for an optimal design of a CMOS operational amplifiers. A multiobjective evolutionary design methodology is used in [6] for the design of a 7-block hierarchical decomposition of a complex high-speed Delta/Sigma A/D modulator.

The purpose of this paper is to develop and implement an algorithm for the design optimization of analog circuits, based on a multiobjective GA. The underlying MOO engine makes use of a genetic algorithm where the fitness of each individual in the population is computed using a Pareto multiobjective raking. Our proposed design method provides a set of optimal solutions (Pareto front) giving the possibility for the designer to select the final one in accordance to further preferences or constrains. The objective functions are formulated using fuzzy sets, while the evaluation of the current design is performed by means of neuro-fuzzy models of circuit performances.

2 Multiobjective Optimization

Multiobjective optimization is concerned with the minimization of a vector of objectives $f(x)$ that may be subject to a number of constrains or bounds:

$$\text{Find a vector } x \text{ that} \quad \text{minimizes} \quad \{f_1(x), f_2(x), \ldots, f_n(x)\} \text{ subject to:}$$
$$g_j(x) \le 0, \quad j = 1,\ldots,m \;; \quad h_q(x) = 0, \quad q = 1,\ldots,p\,; \quad x_l \le x \le x_u \,. \tag{1}$$

where $g_j(x) \le 0$ are inequality constrains, $h_q(x) = 0$ are equality constraints, and x_l and x_u are the lower and upper bounds of the variable vector x.

Because $f(x)$ is a vector, if any of its components are competing, there is no unique solution to this problem. Instead, the concept of noninferiority (also called Pareto optimality) must be used to characterize the objectives [7].

Following the well known concept of Pareto dominance, in the case of all objective functions minimization, an objective vector $f(x^1)$ is said to dominate another objective vector $f(x^2)$, if no component of $f(x^1)$ is greater that the corresponding component of $f(x^2)$ and at least one component is smaller. Accordingly, we can say that a solution x^1 is better than another solution x^2, i.e., x^1 dominates x^2 if $f(x^1)$ dominates $f(x^2)$ [3].

A solution x^* is said to be Pareto optimal, or a nondominated solution for a multiobjective optimization problem (all objectives minimization) if and only if there is no x such that

$$\begin{array}{l} f_i(x) \le f_i(x^*) \text{ for } \forall i = 1,\dots,n \\ \exists j \text{ for that } f_j(x) < f_j(x^*) \end{array} . \tag{2}$$

A nondominated solution is one in which an improvement in one objective requires the degradation of another. Optimal solution, i.e., solution nondominated by any other solution, may be mapped to different objective vectors. In other words, several optimal objective vectors representing different trade-offs between the objectives may exist. The set of optimal solutions is usually denoted as Pareto set, and its image in the objective space is denoted as Pareto front. With many multiobjective optimization problems, knowledge about this set helps the decision maker (circuit designer) in choosing the best compromise solution. In the following, we will assume that the goal of optimization is to find or approximate the Pareto front.

3 The Proposed Algorithm

Design optimization of an electronic circuit is a technique used to find the design parameter values in such a way that the final circuit performances meet the design requirements as close as possible.

To solve this multiobjective optimization problem our approaches consider a genetic algorithm to find or to approximate the Pareto set. Formulating the design objectives for a real design is not always a simple task. The designers can usually accept a certain degree of fulfillment of the design objectives.

In this paper, fuzzy sets are used to define the objective functions [8]. We will associate with each requirement one or two fuzzy sets whose membership functions will represent the corresponding fuzzy objective functions. The fuzzy objective functions became:

$$\mu_k(f_k(x)): \; D_{f_k} \to [0,1] \; . \tag{3}$$

where D_{f_k} is the range of possible values for $f_k(x)$, x is the vector of the design parameters, and f_k is the k^{th} performance function. $\mu_k(f_k(x))$ indicates the error degree in accomplishing the k^{th} requirement. A value μ_k=0 means full achievement of fuzzy objective, while a value μ_k=1 means that the fuzzy objective is not achieved at all. The formulation of the multiobjective optimization problem now becomes:

$$\text{Find } x \text{ that minimizes } \; \{\mu_1(f_1(x)), \mu_2(f_2(x)), \dots, \mu_n(f_n(x))\} \; . \tag{4}$$

where n is the number of requirements.

The performance functions used in our algorithm consist of neuro-fuzzy models of circuit performances. These neuro-fuzzy models (first order Takagi-Sugeno neuro-fuzzy systems [9]) are built up based on input-output data sets using ANFIS [10].

The heart of the whole algorithm is the optimization engine. A genetic algorithm (GA) is responsible for the exploration of the solution space in the quest for the optimal solutions. Generally, the best individuals of any population tend to reproduce and survive, thus improving successive generations [11]. However, inferior individuals

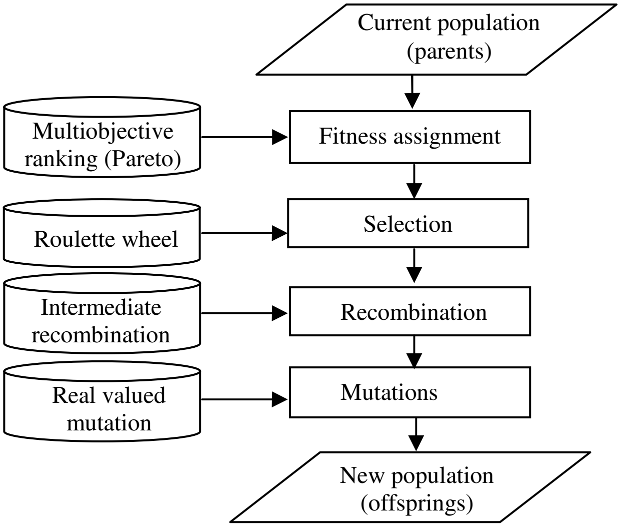

Current population (parents)
Multiobjective ranking (Pareto)
Fitness assignment
Roulette wheel
Selection
Intermediate recombination
Recombination
Real valued mutation
Mutations
New population (offsprings)

Fig. 1. Genetic algorithm multiobjective optimization

can, by chance, survive and reproduce. In our case, the individuals consist of different versions (same topology, but different parameter values) which can evolve until a set of optimal solutions is reached (in terms of requirements accomplishment). The underlying procedure of our GA for multiobjective optimization is presented in Fig.1.

The evolution of the population starts with a random at uniform initialization. Each individual in the current population receives a reproduction probability depending on its own objective function values and the objective function values for all other individuals in the current population.

As long as we are concerned with multiobjective optimization, for the fitness assignment the multiobjective ranking is used, in order to evaluate the quality of an individual. Each individual within a population receives a rank according to its quality. All solutions that are found during optimization and are not dominated by a different solution constitute the Pareto optimal solutions set. All these nondominated solutions will receive a maximum value for their rank.

$$Rank_{max}=N_{Ind}-1\ . \tag{5}$$

where N_{Ind} represents the number of individuals.

For all the other solution the rank is computed using the relation:

$$Rank=Rank_{max}-N_{Dominating}\ . \tag{6}$$

where $N_{Dominating}$ is the number of individuals dominating the individual under consideration. All nondominated solutions have a high selection probability, while the dominated solutions have a lower selection probability, decreasing with the number of dominating individuals.

For the selection, our approach uses the roulette-wheel method. Even if this method is the simplest selection scheme it provides good results, without significant loss of population diversity, when it is used in conjunction with a rank-based fitness assignment (as is the case in this paper), instead of proportional fitness assignment. For each individual *j* a selection probability is computed as:

$$Selection\_probability_j = \frac{Rank_j}{\sum_{i=1}^{N} Rank(i)} \ . \tag{7}$$

where *N* is the number of individuals.

The individuals are mapped to contiguous segments of a line, such that each individual's segment is equal in size to its selection probability. A uniformly distributed random number is generated and the individual whose segment spans the random number is selected. The process is repeated until the desired number of individuals is obtained (called mating population).

Recombination produces new individuals by combining the information contained in two or more parents. For our real valued variables the intermediate recombination method was chosen. Offsprings are produced according to the rule [12]:

$$Var_j^O = a_j Var_j^{P1} + (1 - a_j) Var_j^{P2}, \quad j = 1,\ 2,\ ...,\ Nvar \ . \tag{8}$$

where Var_j^O represent the j^{th} variable of the offspring, Var_j^{P1} represent the j^{th} variable of the first parent, and Var_j^{P2} represent the j^{th} variable of the second parent. The scaling factor a_j is chosen uniformly at random over an interval $[-d, 1+d]$, for each variable. A value of $d = 0.25$ ensures that the variable area of offspring is the same as variable area spanned by the variables of the parents [12].

By mutation, individuals are randomly altered. In [13] it is shown that a mutation rate of 1/*m* (*m*: number of variables of an individual) produced good results for a wide variety of test functions. That means that per mutation only one variable per individual is changed/mutated. Such an operator [12] was considered here:

$$\begin{aligned}
&Var_j^{Mut} = Var_j + s_j r_j a_j, \quad j = 1,\ 2,\ ...,\ m \\
&s_j \in \{-1,\ +1\} \text{ uniform at random} \\
&r_j = r \cdot domain_j, \ \ r \text{ - the mutation range (standard 10\%)} \\
&a_j = 2^{-uk},\ u \in [0,\ 1] \text{ uniform at random;} \quad k \text{ - mutation precision}
\end{aligned} \tag{9}$$

In the above equations, $domain_j$ represent the domain of the variable Var_j and *k* parameter defines indirectly the minimal step-size possible and the distribution of mutation steps in the mutation range. Typical values for *k* are $k \in \{4,5,...,20\}$ [12].

Our GA uses the pure reinsertion scheme: produce as many offsprings as parents and replace all parents with offsprings. Every individual lives one generation only.

4 Experimental Results

Our optimization algorithm is developed in the Matlab. It accepts three types of requirements “greater than”, “equal” and “smaller than” for each design requirement.

We used our algorithm to design a CMOS simple transconductance amplifier (SOTA). Due to the lack of space, is not given here, but it can be found easily in the literature [14]. The circuit is designed for a set of three requirements: voltage gain – *Av*, gain-bandwidth product – *GBW* and common mode rejection ratio – *CMRR*, using four design parameters: three transistor sizes $(W/L)_{12}$, $(W/L)_{34}$, $(W/L)_{56}$, and a biasing current *Ib*. All design parameters have lower and upper bounds, determined so that the transistors in the circuit will remain in their active region regardless the combination of parameter values. These bounds are: *LB*=[20, 0.5, 0.75, 1] and *HB*=[70, 4, 7.5, 100]. The values of GA parameters used in our experimentations are: *d*=0.1, *r*=0.1, *k*=18, *m*=4 and a recombination rate of 1.

The design optimization of SOTA is first illustrated for a set of “equal” requirements as they are presented in Table 1. The optimization was run for a population of 400 individuals for 1000 generations (iterations).

Table 1. Performances and objective functions for “equal” type requirements

Requirements		*Av*	*GBW*	*CMRR*
		=40	=5000 [kHz]	=500000
Indiv.1	Performances	40.46	4891	487418
	Obj. function	*0.0008*	*0.0012*	*0.0013*
Indiv.2	Performances	40.04	4613	472018
	Obj. function	*7.3897e-6*	*0.0156*	*0.0063*
Indiv.3	Performances	40.81	4980	466923
	Obj. function	*0.0025*	*4.09479e-5*	*0.0088*
Indiv.4	Performances	41.37	4716	500158
	Obj. function	*0.0071*	*0.0084*	*2.96298e-5*

At the end of the optimization our algorithm found 34 individuals on the Pareto front. Due to the lack of space we present the performances and the values of objective functions for four of them in Table 1. Indiv.1 was selected as the one with minimum average objective function (0.0011). Indiv.2 is the better one from the point of view of *Av* requirement, meaning that it has a minimum value of the objective function for *Av* in the entire Pareto set (7.3897e-6). Indiv.3 is the one having the minimum objective function for *GBW* requirement in the final Pareto set (4.09479e-5). From the point of view of *CMRR* the best individual is Indiv.4 whose objective function is 2.96298e-5. Each individual constitutes a feasible design solution, the final decision being made by the circuit designer.

The individuals (values of the design parameters) are presented in Table 2. The dynamical behavior of our optimization algorithm is presented in Fig.2. The quality of the entire population is improved generation by generation especially at the beginning of the optimization. The average of the objective functions in the entire population decreases continuously from an initial value of 0.4109 down to 0.0046.

Table 2. Individuals for "equal" type requirements

	Design parameters			
	$(W/L)_{12}$	$(W/L)_{34}$	$(W/L)_{56}$	*Ib*[μA]
Indiv.1	62.3	2.6	7.3	100
Indiv.2	59	2.5	7.4	94.1
Indiv.3	63.1	2.7	7.3	94.1
Indiv.4	57.4	2.7	7.1	94.4

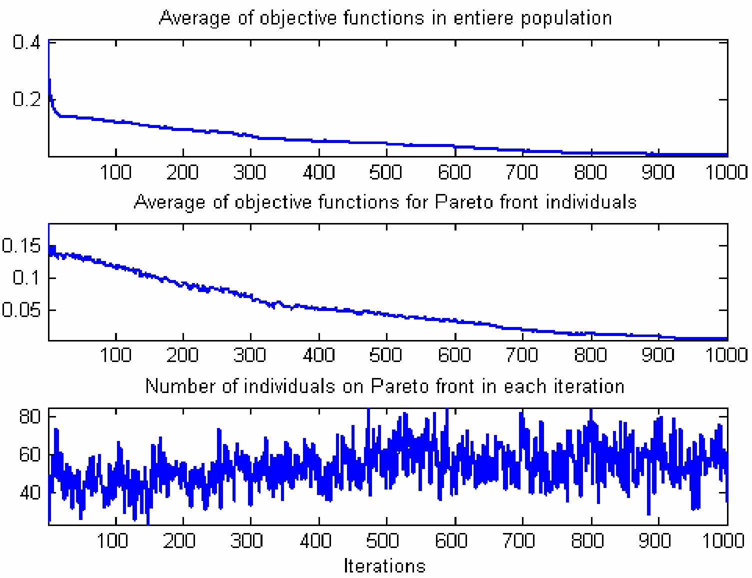


Fig. 2. Dynamic behavior of the optimization algorithm

The quality of the Pareto front is improved during optimization in a slightly oscillating manner, meaning that not always a new Pareto front is better than the previous one. The average of the objective functions in the Pareto front decreases from an initial value of 0.18254 down to a final value of 0.0038. The dimension of the Pareto front varies during the evolution (Fig.2), with a minimum of 23 individuals in generation 149 and a maximum of 85 individuals in generations 474 and 588.

The design optimization of SOTA is then illustrated for a set of "greater than" requirements: *Av*>50, *GBW*>3000[KHz], *CMRR*>1200000.The optimization was run for a population of 100 individuals for only 50 generations. In the population evolution there was only 1 individual on the Pareto front up to the iteration 41. From that point forward, the number of individuals on the Pareto front was increased continuously up to the final value of 97 individuals (out of 100). The performances of five individuals and the corresponding individuals from the final Pareto front are presented in Table 3; all design requirements being accomplished. The individuals are quite similar to each other, a possible interpretation being that during the evolution some diversity of

Table 3. Optimization results for "greater than" type requirements

Performances			Individuals			
Av	*GBW*	*CMRR*	$(W/L)_{12}$	$(W/L)_{34}$	$(W/L)_{56}$	*Ib*[μA]
59.83	3369.16	1252521.98	20.20	3.60	5.70	84.20
59.68	3389.05	1234074.67	20.30	3.60	5.80	82.50
60.90	3468.18	1242884.89	20.20	3.80	5.70	82.10
59.84	3369.10	1220738.45	20.20	3.60	5.60	82.30
58.98	3363.40	1211318.90	20.40	3.50	5.90	81.20

population was lost. The genetic algorithm can be improved if some condition to preserve the population diversity is introduced.

5 Conclusions

A method to design analog circuits using a GA-based multiobjective optimization was presented in this paper. The method uses a multiobjective ranking procedure to compute the fitness of individuals. The algorithm was used to design a CMOS amplifier for different sets of requirements. The algorithm always produces a set of Pareto optimal solutions, regardless the type of requirements ("equal" or "greater than"). The algorithm is an efficient one, the individuals in the Pareto front being permanently improved by evolution.

Further research work should be performed to improve the algorithm by introducing an elitist solution and to maintain population diversity.

References

[1] Greenwood, G.W., Tyrrell, A.M.: Introduction to Evolvable Hardware. IEEE Press Series on Computational Intelligence. Wiley&Sons Inc., Los Alamitos (2007)

[2] Nicosia, G., Rinaudo, S., Sciacca, E.: An Evolutionary Algorithm-based Approach to Robust Analog Circuit Design using Constrained Multi-objective Optimization. Knowledge-based Systems 21(3), 175–183 (2008)

[3] Zitzler, E., Laumanns, M., Bleuler, S.: A Tutorial on Evolutionary Multiobjective Optimization. In: Proceedings of the Workshop on Multiple Objective Metaheuristics, pp. 3–38. Springer, Heidelberg (2004)

[4] Yaser, M.A.K., Badar, K., Faisal, T.: Multiobjective Optimization Tool for a Free Structure Analog Circuits Design Using Genetic Algorithms and Incorporating Parasitics. In: Proc. of the Conference Companion on Genetic and Evolutionary computation, pp. 2527–2534 (2007) ISBN 978-1-59593-698-1

[5] Goh, C., Li, Y.: Multi-objective Synthesis of CMOS Operational Amplifiers using a Hybrid Genetic Algorithm. In: Proc.of the 4th Asia-Pacific Conf. on Simulated Evolution and Learning, pp. 214–218 (2002)

[6] Eeckelaert, T., Schoofs, R., Gielen, G., Steyaert, M., Sansen, W.: Hierarchical Bottom-up Analog Optimization Methodology Validated by a Delta-Sigma A/D Converter Design for the 802.11a/b/g standard. In: Design Automation Conf., 43rd ACM/IEEE, pp. 25–30 (2006)

[7] Boyd, S., Vandeberghe, L.: Introduction to Convex Optimization with Engineering Application. Stanford University (1999)

[8] Oltean, G.: FADO - A CAD Tool for Analog Modules. In: Proc. of the International Conference on Computer as a Tool, EUROCON 2005, Belgrade, pp. 515–518 (2005) ISBN 1-4244-0050-3, IEEE catalog number: 05EX1255C

[9] Oltean, G.: Fuzzy Logic Toolbox 2, User’s Guide, on line version, The Math Works, Inc. (2007)

[10] Jang, R.J.-S.: ANFIS, Adaptive-Network-Based Fuzzy Inference System. IEEE Transaction on System, Man, and Cybernetics 23(3), 665–685 (1993)

[11] Cecilia, R., Machado, J.A.T., Cunha, J.B., Pires, E.J.S.: Evolutionary Computation in the Design of Logic Circuits. In: IEEE International Conference on Systems, Man and Cybernetics, pp. 1664–1669 (2007)

[12] Pohlheim, H.: GEATbx Introduction to Evolutionary Algorithms: Overview, Methods and Operators, version 3.7 (November 2005), `http://www.geatbx.com/`

[13] Mühlenbein, H., Schlierkamp-Voosen, D.: Predictive models for the breeder genetic algorithm in continuous parameter optimization. In: Evolutionary Computation, vol. 1(1), pp. 25–49 (1993) ISSN 1063-6560

[14] Oltean, G.: Fuzzy Techniques in Analog Circuit Design. WSEAS Transactions on Circuits and Systems 7(5), 402–415 (2008)

Optimization of Reconfigurable Multi-core SOCs for Multi-standard Applications

Ali Ahmadinia[1], Tughrul Arslan[2], and Hernando Fernandez Canque[1]

[1] School of Engineering and Computing, Glasgow Caledonian University, UK
[2] School of Engineering and Electronics, University of Edinburgh, UK

Abstract. Today there is a need for high performance chips that can provide very low power consumption, yet can operate over a number of application standards, such as operating a number of telecommunication standards depending on which country the device is in. This paper presents a new framework to enable the design of flexible systems by incorporating different range of reconfigurability in an embedded platform within an SOC design automatically. The SOC design automation involves identifying the best architectural features for the SOC platform, the configuration setting of reconfigurable cores, the type of interconnection schemes, their associated parameters such as data bandwidth, and placement of embedded cores in the communication infrastructures. For this optimization problem, a two-stage multi-objective optimization algorithm is presented. A multi-standard wireless telecommunication protocol is used to demonstrate our optimized designs in terms of area, power and performance.

1 Introduction

The current design automation technology has not been able to match the advance in System-On-Chip technology [1], therefore the ability to use the increasing number of gates effectively has decreased. Existing design methodologies are restricted and mainly based on verification. This paper presents a new design tool to automatically create a digital system directly from a description of the application described in software. In addition, an intelligent multi-objective optimization algorithm is provided to tune the architecture for optimal results under different configurations.

Recently numerous reconfigurable architectures have emerged that can be embedded within an SOC platform. Custom reconfigurable embedded cores can be reconfigured with a small set of configuration bits rather than reconfiguring millions of switch boxes like FPGAs. These provide the advantage of high performance as well as flexibility to future upgrades and dynamic reconfiguration. Reconfigurable SOC architecture incorporates both fixed and newly emerging custom reconfigurable cores. To our best knowledge, there is no existing approach for concurrent optimization of system-level placement of embedded cores and interconnection topology in such reconfigurable SOC architectures. Custom reconfigurable cores can provide multiple standard applications in a single chip, i.e. the resultant architecture should be optimal in different configuration of reconfigurable cores. These optimizations will minimize the overall power consumption and resource area utilization and maximize the whole system's throughput.

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 515–522, 2009.

2 Related Work

There has been considerable effort in designing tools that enable a designer to measure various performance metrics of System-on-Chips. Platune framework [2] tunes performance and power consumption of SOC platforms. Platune is used to simulate an embedded application that is mapped onto the SOC platform and output performance and power metrics for any configuration of the SOC platform. It proposes a space exploration algorithm based on the dependency between parameters. The basic idea is to cluster dependent parameters and then carry out an exhaustive exploration within these clusters. If the size of these clusters increases too much, due to great dependency between the parameters, the approach becomes a purely exhaustive search, with a consequent loss of efficiency.

In ARTS [3], a multi-objective genetic algorithm is proposed to optimize mapping a set of task graphs onto a heterogeneous multiprocessor platform. The objective is to meet all real-time deadlines subject to minimizing system cost and power consumption, while staying within bounds on local memory and interface buffer sizes.

Ascia et al. [4] propose a strategy for exploration of the architectural parameters of the processor, memory, and bus making up a parameterized SOC platform with tight power consumption and performance constraints. It uses multi-objective genetic algorithms as optimization technique for DSE.

Most of the tools outlined so far are designed for tuning of design parameters and subsystems toward a single fixed optimal SOC solution. Our proposed framework (ReCAD) extends such work further by allowing for power, area and throughput analysis of an entire parameterized reconfigurable SOC platform, to achieve a number of optimal results (for each scenario or standard) on the reconfigurable SOC platform within its range of configuration. Furthermore, while the earlier work has either focused on simulation of a user-selected configuration or design space exploration, ReCAD allows for automatic search and exploration of Pareto-optimal configurations with the ability of transaction level simulation of whole system for more accurate performance evaluation and speeding up system verification.

3 ReCAD Framework

The main aim of developed ReCAD tool is to automate the development of multi-core architectures, incorporating custom reconfigurable components, conventional RISC based processors, hardware accelerators and memory blocks etc. The components are added in the built-in library. Depending on the application required, components can be chosen to be integrated. In order to get the optimized system, the tool selects the communication media and embedded cores from the library of components with characteristics that satisfy the user constraints.

In order to verify the functionality of the many different configurations generated by platform transformations, there is a need for a fast simulation. To facilitate the ease of simulation, the most successful approaches are based on SystemC [6].

For power estimation, a state-based power model integrated with SystemC is used. Throughput estimations are mainly based on the SystemC simulations. A design space exploration will be used to search possible interconnection schemes with optimal placement of components in the communication medium to meet power and performance constraint of each application scenario.

4 Architectural Tuning for Multi-standard Applications

This section discusses the development of what we term the architectural tuning engine, which deals with:

a. Identifying the best architectural features for the SOC platform including embedded core types and interconnection schemes.
b. Identifying best parameter sets associated with each entity including configuration settings and bus bandwidth.
c. Determining the optimal core placement in the communication medium.

The above tasks should be carried out with considering the multi-scenario application to meet all scenarios' (standards') constraints. To identify the optimal set of embedded cores, interconnection topology and its associated parameters, a mapping of components to the communication medium is necessary. The power and throughput metrics may vary with different placement of modules in the interconnection scheme. For this reason, objectives should be examined against each feasible placement to identify the optimal solution which its performance results meet all required scenarios' specifications.

Like similar approaches, the ReCAD framework includes RTL implementation of each component of its library. Through one-off RTL synthesis of components, their resource area utilizations are calculated. In ReCAD tool, SystemC model for each SOC platform component as object has been developed which then could be used to model the complete SOC platform as communicating objects.

A SystemC based power model is developed for measurement of component's switching activities to obtain power estimation more accurately [5]. It is important to notice that gate-level simulation (for power analysis) has to be performed only once for each component. Therefore, when executing the complete system level model of a particular configuration, fast and sufficiently accurate power estimates will be obtained. Moreover, the SystemC models of components enable us to estimate overall throughput more accurately with a fast SystemC simulation compared to RTL based estimations. We assumed that power and throughput estimations of communication media depend on the transaction size of data between two ports. Since, the distance and type of two communicating ports have impact on the throughput and power consumption of the system, the transaction size is not the only major factor for performance estimation, especially in hybrid interconnections and NoCs. For this reason, the proposed SystemC power simulation is used for different traffic patterns between each two ports of the interconnection schemes. Thus, an approximate throughput and power consumption of the interconnection medium will be obtained for each communication scenario. With this information, we are able to estimate the overall area, power, and throughput of a complete reconfigurable SOC platform in different configuration modes to examine whether the estimated reconfigurable SOC solution meets the required constraints of each application scenario.

So far, we have described the simulation model and power analysis techniques used in the ReCAD tool. In the next section, we formulate the exploration problem and outline the algorithms used for performing it automatically.

5 Design Space Exploration

5.1 Problem Definition

Let *S* be a parameterized reconfigurable SOC platform with *n* parameters. The generic parameter p_i, *i*=1, 2, … , *n*, can take any value in the set V_i. A complete assignment of values to all the parameters is a configuration. The problem is to efficiently, compute the Pareto-optimal configurations, with respect to power, area and throughput, for a multi-scenario application executing on the reconfigurable SOC platform. In our problem, a configuration C_i is Pareto-optimal if no other configuration C_j has better power as well as area and throughput than C_i.

Our SOC platform is composed of numerous embedded cores and interconnection schemes. The type of embedded cores and interconnection scheme parameters need to be identified and fixed for the application. However, configuration settings of reconfigurable embedded cores provide flexibility to the SOC platform in order to execute multiple standard applications. In contrast with traditional design space exploration problem that an optimal fixed configuration should be achieved for a fixed application, here a reconfigurable SOC platform for a multi-standard application should be provided.

Therefore, we adapt traditional DSE problem to meet our requirement. We define two sets of parameters for optimization. The first set includes the configuration setting of reconfigurable cores and the second one includes the rest of parameters. Now, we use a two-stage design space exploration: in the first stage the configuration space should find Pareto-optimal solutions for the application scenario with the $min(Objective_1)$ among all application scenarios, for example a standard with maximum power consumption constraint. In this stage all reconfigurable cores are set to their largest configuration size. Then in the second stage, there would be a set of optimal solutions that should explored by only the configuration setting parameter to meet constraints of each scenario, which can be explored by heuristics such as hill climbing algorithm.

5.2 Multi-objective Optimization

To obtain an optimized solution for the exploration problem, first we need to define a DSE strategy that will give a good approximation of the Pareto-optimal front for a SOC platform *S* and an application *A*, simulating as few configurations as possible. The search for optimal configurations is a question of multi-objective optimization, where some of the objectives conflict with others, for example, performance and power consumption. Although this causes a considerable increase in the complexity of DSE strategies, it has the advantage of offering the SOC designer not one but a set of optimal configurations (Pareto-optimal set) from which he can choose the one that represents the best tradeoff in relation to the set of constraints has to be met. There are two main approaches for DSE of SOC platform. The first, GA, uses Genetic Algorithms as the optimization engine. A configuration is mapped onto a chromosome and a population of configurations is made to evolve until it converges on the Pareto-optimal set.

The second approach is the interdependency model proposed in [2], which tries to prune the configuration space. They have used a graph model to capture the parameter interdependencies. Such a graph is constructed with its nodes representing parameters and edges representing interdependencies between parameters.

This algorithm consists of two phases. The first phase performs a local search for Pareto-optimal configurations. The phase performs clustering of interdependent nodes in the graph. This is the same problem as finding strongly connected components of a graph (e.g., a depth first search can be used to accomplish this). The second phase iteratively expands the local search to discover global Pareto-optimal configurations. The stage combines pairs of clusters into a single cluster and computes Pareto-optimal configurations within it. Then, it limits the space of this new cluster to the Pareto-optimal configurations only. This procedure is repeated until all the clusters have been merged and a single cluster remains. The Pareto-optimal configurations within this last cluster represent Pareto-optimal configurations of the entire configuration space.

5.3 Genetic Algorithm

The approach we propose for exploration of the configuration space of a parameterized reconfigurable SOC uses both inter-dependency and GA algorithms. For GA approach, we chose SPEA2 [6], which is very effective in sampling from along the entire Pareto-optimal front and distributing the solutions generated over the trade-off surface. The representation of a configuration can be mapped on a chromosome whose genes define the parameters of the system. The gene coding the parameter P_i can only take the values belonging to the set V_i. The chromosome of the GA will then be defined with as many genes as there are free parameters and each gene will be coded according to the set of values it can take (Fig. 1). For each objective to be optimized, it is necessary to define the respective measurement functions. These functions, which we will call objective functions, frequently represent cost functions to be minimized (area, power, and throughput).

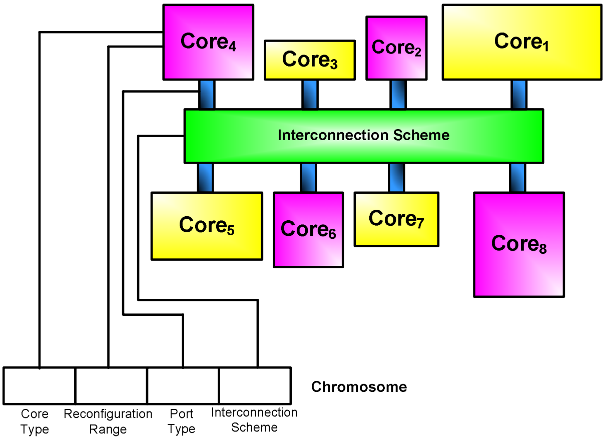


Fig.1. Representation of a configuration as a chromosome

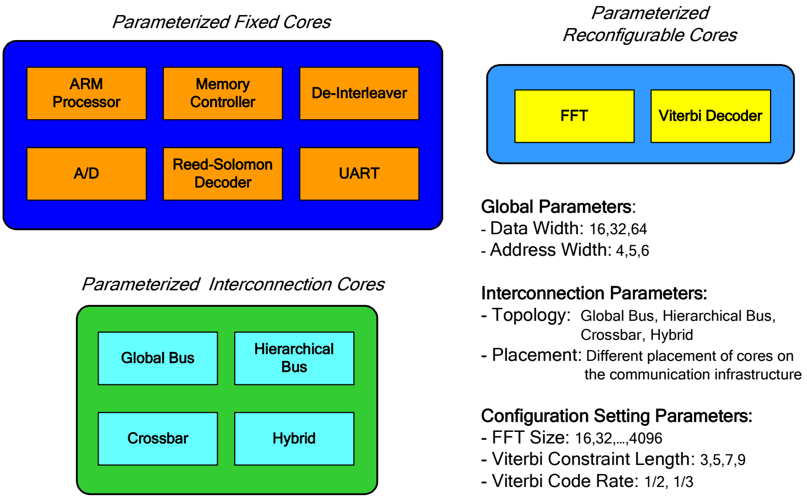
6 Result Analysis

The proposed tuning algorithm has been applied to a multi-standard telecommunication application which handles different standards: WiMAX, WLAN, GSM, and 3G-CDMA with embedded custom reconfigurable cores of FFT and Viterbi decoder.

Fig. 2 shows the parameterized embedded cores targeting multiple standard wireless SOC devices used in our experiments. In our parameterized reconfigurable SOC architecture, there are two global parameters (data and address-width) to parameterize data bandwidth of embedded cores and interconnection schemes. Unlike other approaches, we use two parameters for interconnection cores: topology and placement. Topology parameter chooses the topology of interconnection structure, and Placement parameter determines the placement of embedded cores in the communication infrastructure. To obtain estimation values of power and area for the placement parameter, a set of different traffic patterns has been simulated between each two ports that their average power consumption and throughput give a realistic estimation of the interaction between ports to evaluate the effect of different placements. Moreover, configuration setting parameters should be optimized for a range of configurations to give optimal solution in different wireless standards.

We explored the configuration space for the mentioned application standards. In the first stage, we have optimized our architecture with configuration setting parameters for the 3-G CDMA standard (FFT Size=4096, Constraint Length=9, Code Rate=1/3). In this stage, we have simulated our architecture with two algorithms: the GA one and the interdependency approach [2].

With the set of optimal solutions obtained from the first stage, in the second stage, these solutions are explored by only the configuration setting parameters to meet constraints of each standard with a hill climbing algorithm. Results are summarized in



Global Parameters:
- Data Width: 16,32,64
- Address Width: 4,5,6

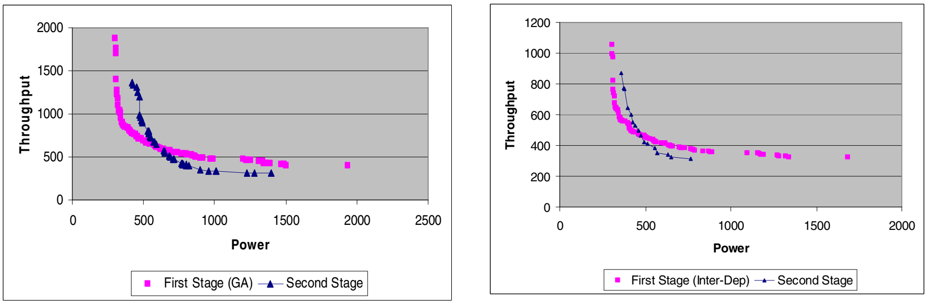
Interconnection Parameters:
- Topology: Global Bus, Hierarchical Bus, Crossbar, Hybrid
- Placement: Different placement of cores on the communication infrastructure

Configuration Setting Parameters:
- FFT Size: 16,32,...,4096
- Viterbi Constraint Length: 3,5,7,9
- Viterbi Code Rate: 1/2, 1/3

Fig. 2. Parameterized cores for optimization of a reconfigurable SOC architecture for multiple wireless standard applications

Table 1. Two Stage Optimization Results

Stage	GA BASED					Inter-Dependency Based				
	Config. Space	Pareto Optimal	Throughput Trade-off	Power Trade-off	Area Trade-off	Config. Space	Pareto Optimal	Throughput Trade-off	Power Trade-off	Area Trade-off
1	967670	396	7.4	8.3	9.1	967670	743	7.4	8.3	9.1
2	28512	26	6.3	6.4	4.5	53496	43	5.5	5.5	4



GA algorithm in the first part

Inter-dependency algorithm in the first part

Fig. 3. Power/Throughput Tradeoff in the two stage optimization with different algorithms in the first stage and hill climbing heuristics in the second stage

Table 1. The power, throughput, and area trade-offs of the Pareto-optimal configurations for all 4 application standards are presented. In the second stage, average trade-offs decreased with the interdependency algorithms [2] due to its effective configuration pruning algorithm.

Fig. 3 presents the power/throughput tradeoff results in our two stage optimization with the two different algorithms used in the first stage: GA and Inter-dependency approaches. The simulation times with the GA algorithm were in the range of few seconds, whereas the Inter-dependency approach needed hours to compute the optimal solutions. On the other hand, the Inter-dependency approach was more effective in pruning the configuration space. The simulation times of our second stage were in the range of minutes, since it optimized a small configuration space with a few configuration setting parameters.

7 Conclusion

This paper has presented a tool for automatic creation of digital systems directly from a description of the application described in software. Further to this, an intelligent multi-objective optimization algorithm has been specially tailored to tune the architecture for optimal results with different configuration. This proposed novel approach involves concurrent optimization of system-level placement of embedded cores and interconnection topology within a custom reconfigurable SOC architecture for a multi-scenario application. Custom reconfigurable cores can provide multiple standard applications in a single chip, i.e. the resultant architecture is optimized in different configuration of reconfigurable cores. These optimizations minimize the overall power consumption, resource area utilization and maximize the whole system's throughput. This approach has been applied to a wireless multi-standard application, and the results demonstrate the feasibility of our proposed approach.

References

1. International Technology Roadmap for Semiconductors (ITRS), 2005 edn
2. Givargis, T., Vahid, F.: Platune: A Tuning Framework for System-on-a-Chip Platforms. IEEE Tran. on Computer Aided Design 21(11), 1317–1327 (2002)
3. Madsen, J., Stidsen, T.K., Kjaerulf, P., Mahadevan, S.: Multi-Objective Design Space Exploration of Embedded System Platforms. In: Conference on Distributed and Parallel Embedded Systems (DIPES), Braga, Portugal, October 11-13, 2006, pp. 185–194 (2006)
4. Ascia, G., Catania, V., Palesi, M.: A GA based design space exploration framework for parameterized system-on-a-chip platforms. IEEE Transactions on Evolutionary Computation 8(4), 329–346 (2004)
5. Ahmadinia, A., Ahmad, B., Arslan, T.: Efficient High-Level Power Estimation for Multi-Standard Wireless Systems. In: Proceedings of IEEE Computer Society Annual Symposium on VLSI (ISVLSI), Montpellier, France, April 7-9, 2008, pp. 275–280 (2008)
6. Zitzler, E., Laumanns, M., Thiele, L.: SPEA2: Improving the performance of the strength pareto evolutionary algorithm. In: Evolutionary Methods for Design, Optimization and Control with Applications to Industrial Problems, Athens, Greece, pp. 95–100 (2001)

Vocabulary Learning Environment with Collaborative Filtering for Support of Self-regulated Learning

Masanori Yamada[1], Satoshi Kitamura[1], Shiori Miyahara[2], and Yuhei Yamauchi[1]

[1] Interfaculty Initiative in Information Studies, the University of Tokyo,
7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-0033, Japan
{masanori-y,satkit,yamauchi}@iii.u-tokyo.ac.jp
[2] Consortium for Renovating Education of the Future, the University of Tokyo,
7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-0033, Japan
miyahara@coref.u-tokyo.ac.jp

Abstract. This study elucidates issues related to using online vocabulary learning environments with collaborative filtering and functions for cognitive and social learning support in learner-centered learning, which requires learners to be self-regulated learners. The developed system provides learners with a vocabulary learning environment using online news as a test installation of functions. The system recommends news to each learner using a collaborative filtering algorithm. The system helps learners to use cognitive and social learning strategies such as underlining, along with a word-meaning display based on the learner's vocabulary proficiency level. We investigated effects of the system on perceived usefulness and learning performance as a formative evaluation. Learners regarded this system as a useful tool for their language learning overall, but rated several functions low. Confirming the learning performance, the learner's vocabulary proficiency level improved significantly.

Keywords: Language learning, Educational technology, Learning strategies, Aptitude treatment interaction, Collaborative filtering.

1 Introduction

Concomitantly with the advancement of information communication technology, interest in using online learning environments for language learning has grown. Nevertheless, learners must be self-regulated when using learner-centered learning environments without instructors, such as online learning. Self-regulation is an important factor in successive language learning [1]. That is particularly true in Japan, where Japanese people have few opportunities to communicate with foreign language speakers. Consequently, a common topic is how to support online learning activities to foster self-regulated learners.

For self-regulated learners, the effect of the online language learning environment on learning performance is limited because this environment specifically includes self-regulated learners [2]. It is apparently difficult for non-self-regulated learners to continue to learn foreign languages in an online language learning environment. Therefore, learning activities can be supported based on three points of view: cognitive learning strategies, motivation, and social support.

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 523–530, 2009.

First, cognitive learning strategies directly promote understanding of the learning objective. Promotion of the use of cognitive learning strategies is apparently effective for the support of self-regulated learning [3]. However, instruction about learning strategies in a face-to-face classroom environment blended with online learning [4], and socio-affective strategies such as pair learning [5] are mainly conducted because improvement and research related to language learning specifically emphasizes face-to-face instruction. Consequently, it is necessary to develop functions in online language learning environments for the use of cognitive learning strategies.

Secondly, motivation is a central factor in successive learning. Especially, potential problems related to motivation arise in online learning environments. One is connected with the learning materials themselves. In language learning, input information such as learning material has a strong effect on learning performance [6] and motivation [7]. Input means written or spoken information in the target language that the learner can comprehend (e.g. [6], [8]). It is useful to refer to aptitude treatment interaction (ATI) for learning contents and system development. Actually, ATI signifies the interaction effect of learner's features and learning contents on learning performance, i.e., the effect of learning materials depends on the learner's independent features such as the learner's level, learner's interest, and hobbies [9]. Learning support considering ATI is apparently effective on learning performance and on the enhancement of learners' motivation. It is difficult for an online learning environment to provide appropriate learning materials to each learner because teachers must prepare learning materials conforming to each learner's prior knowledge and preferences. The other factor is related to isolation. Learners tend to be isolated in an online learning environment, apart from the face-to-face classroom. In this situation, learners share and receive little information and feedback that is usually effective for promotion of learning from the social affective and learning consciousness perspectives [10]. This lack of feedback inhibits learners' motivation. Communication tools such as Bulletin Board Systems are often added to online learning environments to solve this problem, but few learners use communication tools because of the lack of context in which to use them.This study is intended to develop English news distribution systems for English vocabulary learning, considering the enhancement of self-regulated learning from the viewpoints of self-experience and social assistance, which facilitate self-regulated learning processes [11]. We evaluate this system from the viewpoints of learners' perceived satisfaction and learning performance as a formative evaluation phase.

2 System Development

2.1 System Architecture

This system is a client/server system. Clients include software allowing the support functions of cognitive learning strategies, and communication for social learning support among learners. Client software was developed using Asynchronous JavaScript and XML (AJAX). The server side includes software for English news distribution, in addition to storage of learners' use of the support function of cognitive learning strategies. These functions in software are implemented mainly in JAVA, and partially PHP (news scraping). The server system works on an Apache 2.0 web server with the PHP module, and a JBoss application server 3.2.7. PostgreSQL 8.1.4 was used as a database server.

2.2 System Functions

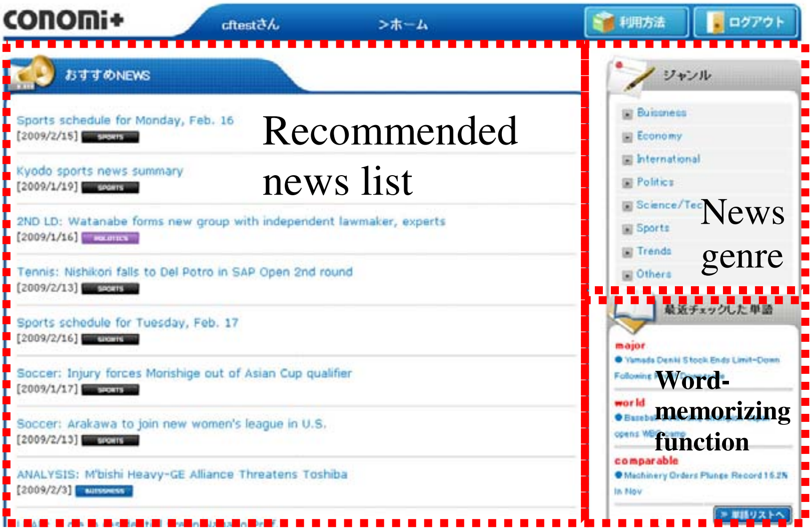
This system includes three components for learners to learn vocabularies through encouragement of self-regulated learning. They are respectively designed for

(1) promotion of use of cognitive learning strategies,
(2) aptitude treatment interaction (ATI), and
(3) support of active interaction and feedback between learners.

The use of cognitive learning strategies such as underlining and motivating with learning materials in which learners have an interest are reportedly effective for fruitful learning experiences [3]. This study first designed and developed underlining, meaning display and word-memorizing functions for support to use cognitive learning strategies (elaborating by highlighting and using dictionary). The underlining function enables a learner to underline sentence(s) or word(s) when the learner finds it important for their learning or interest. Regarding meaning display, this system displays a word meaning when learners put a mouse cursor over an underlined word. The number of words displaying meaning depends on the learner's vocabulary level judged according to The Japan Association of College English Teachers (JACET) vocabulary level, which is a standardized vocabulary level measurement for English learners over college level. The word-memorizing function enables learners to make lists of unknown words. When a learner finds an unknown word, the learner can make such a list by clicking the word. This system records the word and news headers that learners have read.

The second component is used for motivating learners by providing appropriate learning materials, considering ATI. To do so, we implement collaborative filtering for the provision of learning materials along with the learner's interest. Collaborative filtering is an algorithm for the recommendation of information. The system with added collaborative filtering predicts a user's preference based on analyses of a similar user's preference. A collaborative filtering algorithm first finds similar users. Collaborative filtering has various ways to find similar users. An important way to find similar users is to use the correlation rate between users in a preference pattern. Users who have a high correlation rate are regarded as similar users to active users.This study used the "GroupLens" algorithm [12], which is a representative algorithm based on a memory-based approach, for collaborative filtering. This system predicts an active learner's interest in an unread article, calculating Pearson's correlation coefficient using similar users' rating data for read articles. This study is intended to examine the practical use of collaborative filtering in educational settings. Therefore, a simple design and algorithm are preferred for our future research, although many researchers suggest an improved algorithm [13],[14]. Therefore, we used the "GroupLens" algorithm as a test installation.

The last function related to interaction among learners aims to motivate learners to understand the content and active feedback. White (2003) suggests that social learning support such as feedback among learners affects the promotion of self-regulated learning. After the active learner clicks a news reader icon, this system shows news readers each article and an article list of news that readers have already read. Furthermore, learners can comment on each article. This function seems to reduce isolation and encourage readers to read articles, being aware of similar learners. The system interfaces are displayed in Figs. 1 and 2.



Sports schedule for Monday, Feb. 16
[2009/2/15] SPORTS

Kyodo sports news summary
[2009/1/19] SPORTS

2ND LD: Watanabe forms new group with independent lawmaker, experts
[2009/1/16] POLITICS

Tennis: Nishikori falls to Del Potro in SAP Open 2nd round
[2009/2/13] SPORTS

Sports schedule for Tuesday, Feb. 17
[2009/2/16] SPORTS

Soccer: Injury forces Morishige out of Asian Cup qualifier
[2009/1/17] SPORTS

Soccer: Arakawa to join new women's league in U.S.
[2009/2/13] SPORTS

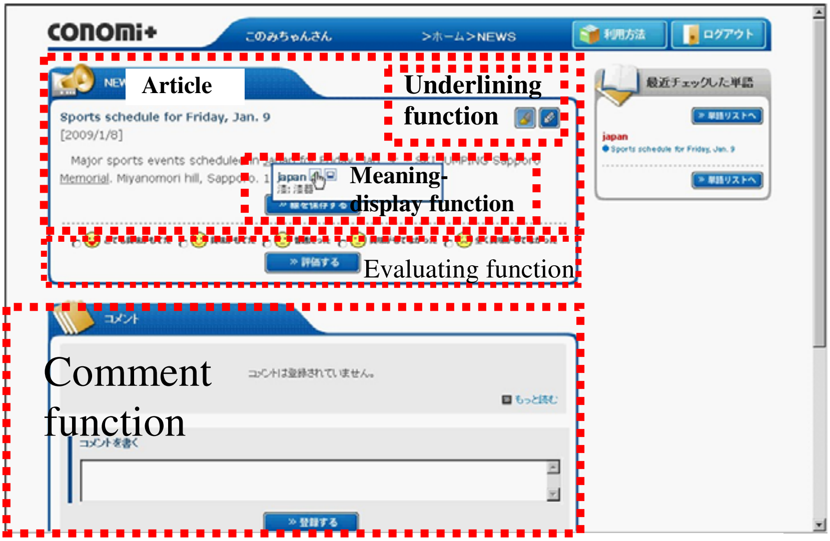
ANALYSIS: M'bishi Heavy-GE Alliance Threatens Toshiba
[2009/2/3] BUSSINESS

major
● Yamada Denki Stock Ends Limit-Down Following

world
● Baseball ... opens WBC camp

comparable
● Machinery Orders Plunge Record 16.2% in Nov

Fig. 1. Interface of top page (displaying recommended news header)



Sports schedule for Friday, Jan. 9
[2009/1/8]

Major sports events schedule in Japan for Friday, Jan. ... SKI JUMPING Sapporo ... Memorial. Miyanomori hill, Sapporo. 1...

japan
訳: 日本

コメントは登録されていません。

japan
● Sports schedule for Friday, Jan. 9

Fig. 2. Word meaning display

3 Methodology

3.1 Subjects

The subjects in this study were 235 university students. All subjects were non-native speakers of English. Of them, 132 subjects were prospective employees of the same company. Of the 235 students, 91 subjects were asked to read and rate the degree of interest in article content every day to support collaborative filtering. Therefore, data of 132 subjects were analyzed in this study for evaluation of this system. The subjects' proficiency in English varied from that of low-level students, who had difficulty in reading short passages, to that of high-level students who had studied English education. All subjects had reached at least a high school standard level in grammar and vocabulary. Because of privacy rules of the company, information about subjects was not provided.

3.2 Procedure

In all, 132 subjects were asked to use this system voluntarily for one month. First, 132 subjects were asked to take a vocabulary test (JACET test) to find a suitable assistance level to display the word meaning, and to evaluate the system's effects. Each subject accesses and uses this system whenever and wherever they want to. An online help file was uploaded on this system because of the lack of opportunity to meet subjects through this term. After one month, subjects were asked to answer questionnaires about their satisfaction with articles and the system's usability. Finally, we asked 132 subjects to take the JACET vocabulary test again.

3.3 Data Collection

This study investigated the contribution of functions to learning performance as formative evaluation. Data were collected in two ways. The first is a questionnaire. All subjects were required to answer a questionnaire after the experiment. The questionnaire asked all subjects to rate the perceived satisfaction with quantities of distributed articles, the degree of interest in article contents, and the degree of perceived helpfulness of each function from a 5-point rating scale.

4 Results

4.1 Perceived Effect and Usability of This System

In all, 56 learners answered the questionnaire. Table 1 presents the number of subjects who rated the degree of perceived effect and usability in each item. The results revealed that this system was apparently effective on their learning and the perceived satisfaction overall. Items related to a learner's interest in news ("This system did not recommend articles in which I have an interest", and "There is no genre that I wanted to read") were rated as effective. Regarding the evaluation of functions, the meaning display function and comment function were rated highly, but the underlining function was not evaluated positively. Learners felt burdened by rating their interest in each article, which plays an important role in predicting learners' interest based on responses of similar learners.

Table 1. Results of the perceived effect of this system

	Strongly agree	Agree	Neither	Disagree	Strongly Disagree
The quantities of recommended news are suitable for me	2	24	11	17	2
I had interest in news recommended by this system	4	19	19	13	1
This system did not recommend articles in which I have an interest	3	16	18	19	0
There was no genre that I wanted to read	2	10	13	25	6
It was troublesome to rate the degree of interest in news	2	19	8	19	8
I did not rate interest in news, but I read it	6	18	6	13	13
It was helpful to use meaning display function for reading the article *1	13	23	10	8	0
The word-memorizing function assisted your learning *2	1	22	18	10	1
The underlining function assisted your learning *3	3	14	21	7	3
Comments from other learners on articles were helpful to comprehend article content	7	21	17	8	3
I learned English, thanks to this system	5	32	14	5	0

*1 Two learners did not use this function.

*2 Four learners did not use this function.

*3 Eight learners did not use this function.

4.2 Learning Performance

We conducted a vocabulary test (JACET test) for evaluation of learning performance. In all, 64 learners took both a pre-test and a post-test. To do so, learners took the JACET test before and after the evaluation term. We calculated each learner's JACET

Table 2. JACET test

	Mean	S.D.		
Pre-test	4.18	2.37	t(63)=3.35	p<0.01
Post-test	4.89	2.42		

level from pre and post JACET test scores; then a paired *t*-test was conducted on the JACET level in the pre-test and post-test. Results of a paired *t*-test are presented in Table 2. Results show that this system improved the learners' vocabulary proficiency significantly.

5 Discussion

Overall, this system seemed to contribute to learners' motivation for learning, and to learning performance. The effect of meaning display, word-memorization, and comment functions on their learning were positively recognized overall, according to the questionnaire results. In particular, the comment function allows learners to comment on the text area in their native language, which facilitates communication between learners and understanding the article. However, two problems were revealed in this formative evaluation. First, collaborative filtering did not work well because of the lack of rating data for articles. The questionnaire results revealed that learners felt that rating the interest in articles was bothersome. This problem can hinder or prevent collaborative filtering. For predictive accuracy, collaborative filtering uses a large amount of rating data to detect learners who are similar to active learners.

Second, the usefulness of underlining functions was rated as less positive. Learners who answered "neither" accounted for a large share of all learners who answered questionnaires. This function was designed based on learning science and psychological theories. However, this function enabled learners to use learning activities. Learners might not use the objective to use these functions.

6 Conclusion and Future Works

This study is intended to develop an English news distribution system with collaborative filtering for English learning, and to evaluate it as formative evaluation. Overall, this system was recognized as a useful tool for English learning. In fact, positive results were revealed in the perceived usefulness of system functions and in learning performance. However, this study revealed several problems. One is the dearth of rating data, which are necessary for effective collaborative filtering. The other is the relation between functions and learning objectives. Learners were apparently aware of using some functions such as underlining and word-memorizing functions. These functions, based on learning theories, should be designed such that learners can understand how to use them to support their own learning.

Future works aimed at realization of suitable learning environments are recommended as follows.

(1): Implementation of a bridge function between the underlining function and comment function

Results of this study indicate that the effects of underlining and word-memorizing functions on learning were unclear. The system must enable learners to be aware of the purpose of using these functions. One way seems to be connection of these functions to comment functions to facilitate social support.

(2): More elaborate analysis

To determine the effect of this system on learning, comparative analyses will be conducted using low-frequency and high-frequency groups, and groups which are less and more familiar with the recommended articles. In the next phase, we will add analyses of self-regulated learning such as the continuance rate and school behaviors.

References

1. Sakai, S.: A study on the relationship between developing learner autonomy and ICT in ELT. Journal of Multimedia Aided Education Research 6(1), 46–56 (2008)
2. White, C.: Language Learning in Distance Education. Cambridge University Press, Cambridge (2003)
3. Garcia, T., Pintrich, P.R.: Regulating motivation and cognition in the classroom: The role of self-schemas and self-regulatory strategies. In: Schunk, D.H., Zimmerman, B.J. (eds.) Self-regulation of learning and performance: Issues and educational applications, pp. 127–153. Lawrence Erlbaum Associates, Hillsdale (1994)
4. Hauck, M.: Metacognitive Knowledge, Metacognitive Strategies, and CALL. In: Egbert, J., Petrie, G.M. (eds.) CALL Research Perspectives, pp. 65–86 (2005)
5. Levy, M.: Computer-Assisted Language Learning – Context and Conceptualization. Oxford University Press, Oxford (1998)
6. Krashen, S.: The input hypothesis: Issues and implications. London House, Harlow (1985)
7. Dörnei, Z.: Motivational strategies in the language classroom. Cambridge University Press, Cambridge (2001)
8. Gass, S., Mackey, A., Pica, T.: The role of input and interaction in second language acquisition. The Modern Language Journal 82(3), 299–305 (1998)
9. Cronbach, L.J.: How can instruction be adapted to individual differences? In: Gagne, R.M. (ed.) Learning and individual differences, pp. 23–39. Charles Merrill, OH, OH (1967)
10. Lou, Y., Dedic, H., Rosenfield, S.: A feedback model and successful e-learning. In: Naidu, S. (ed.) Learning & teaching with technology – Principles and practices (open & flexible learning series), pp. 249–259. Routledge, Falmer, OX, UK (2003)
11. Schunk, D.H., Zimmerman, B.J.: Motivation and Self-Regulated Learning: Theory, Research, and Applications. Routledge, Falmer, London, UK (2007)
12. Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., Riedl, J.: GroupLens: An open architecture for collaborative filtering of netnews. In: Proceedings of the ACM Conference on Computer Supported Cooperative Work, pp. 175–186 (1994)
13. Breese, J.S., Heckerman, D., Kadie, C.: Empirical analysis of predictive algorithms for collaborative filtering. Uncertainty in Artificial Intelligence 14, 43–52 (1998)
14. Herlocker, J.L., Konstan, J.A., Borchers, A., Riedl, J.: An algorithmic framework for performing collaborative filtering. In: Proceedings of 22nd Annual ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 230–237 (1999)

Online Collaboration Support Tools for Project-Based Learning of Embedded Software Design[⋆]

Takashi Yukawa[1], Hirotaka Takahashi[1], Yoshimi Fukumura[1],
Makoto Yamazaki[2], Toshimasa Miyazaki[2], Shohei Yano[2],
Akiko Takeuchi[2], Hajime Miura[3], and Naoki Hasegawa[4]

[1] Nagaoka University of Technology
[2] Nagaoka National College of Technology
[3] Techno Holon Corporation
[4] Niigata Industrial Creation Organization

Abstract. The present paper reports the requirements, design, and learning effects of online collaboration support tools for project-based learning (PBL) applied to the development of embedded software. In this research, the authors created a new program that blends face-to-face classes and e-Learning classes. They also developed a computer-supported collaborative learning environment. In the present paper, the requirements for the collaboration support tools for the learning program are clarified through observation of a real PBL course. Based on this observation, an online repository tool and a unified search tool are proposed and implemented. The online repository tool was applied in a trial course blending face-to-face and online activities. Participants in the trial course completed a questionnaire survey. According to the survey responses, the blended learning program is feasible for PBL of embedded software design, and the online repository tool facilitates collaborative activities between learners and is effective for expanding each learnerfs design ability.

1 Introduction

Project-based learning (PBL) is intended to strengthen studentsf abilities in design, teamwork, and communication through experiences developed in solving practical problems as a team. As an example, PBL has become popular in engineering education because industry requires new university graduates to have engineering design abilities.

However, there are some potential problems. Since group study and meetings are the key to PBL, learners need to spend a great amount of time in the classroom. For part-time students, this can limit their opportunities to attend PBL-based courses. In addition, since learners tackle problems as a team,

[⋆] The present research is supported by a grant through the Strategic Information and Communications R&D Promotion Programme (SCOPE) from the Ministry of Internal Affairs and Communications, Japan.

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 531–538, 2009.

each member of the team must have a certain level of knowledge and skills in the target area. Also, project-based learners must organize their experiences into systematic knowledge and insight to acquire the desired abilities; otherwise, PBL becomes an ineffective and time-consuming process.

To address these problems, the authors launched a project that implements e-Learning technology, a form of Information and Communication Technology (ICT), in PBL [1]. Knowledge and programming skills can be enhanced by self study with conventional lecture-based e-Learning materials. Therefore, collaboration between learners is a significant factor in PBL. With this consideration, this project focused on both the e-Learning program of the collaborative design process and the tools that support the e-Learning program. The present paper reports the requirements and design of online collaboration support tools for PBL in the development of embedded software. Evaluation results on the learning effect of this e-Learning approach and usability of the tools are reported.

2 Background

In this section, a PBL course on the development of embedded software, e-Learning technology, and the concept of the project are briefly introduced.

2.1 A Project-Based Learning Course on Development of Embedded Software

Demand for the development of embedded software has increased because of the widespread use of microprocessors in various types of equipment. However, current training courses on computer software focus primarily on enterprise application software. Therefore, the number of development engineers for embedded software is insufficient. In particular, for Japan to remain globally competitive in the development of industrial products, more embedded systems engineers are needed.

In recent years, a number of training programs on embedded software have been established. In Niigata prefecture, the Niigata Industrial Creation Organization (NICO) and associated organizations have conducted training courses on embedded software since 2006. The NICO program incorporates a PBL course [2].

Potential participants in the training course are assumed to be active engineers who have responsibilities at their workplace. Thus, some of the participants could not easily participate in group training. In a traditional PBL course, since a class cannot be formed unless all members of the team are present, time and/or spatial restrictions become obstacles.

2.2 e-Learning Technology for Collaborative Learning

Research on collaborative learning with the support of ICT has become increasingly active, and so research on Computer Supported Collaborative Learning

(CSCL) has evolved [3,4,5,6]. In a CSCL environment, communication between a teacher and a learner and/or between learners, sharing documents and program codes created through the learning process, and exchange of atmosphere (awareness) among learners are supported by a computer.

2.3 Implementing e-Learning Technology for a Project-Based Learning Course

In PBL, groups of learners work together to achieve the projectfs goal. In this particular study, the learners strengthen abilities for embedded software development through their PBL experiences. The objectives of this research are clarifying how ICT can support each stage of the software development process to improve the effectiveness of learning, and designing a learning course with the support of ICT.

The objectives of the project were accomplished in the following steps:

Step 1. The behaviors of learners are observed in an actual PBL course. Based on the observation, the possibilities of ICT support are considered.

Step 2. Then, based on the findings obtained from the previous step, a PLB-based education program for embedded software development supported by ICT is designed. The supporting computational environment is also developed.

Step 3. Finally, the program and the support system are implemented in an actual course and the effectiveness is evaluated.

The number of educational institutions that incorporate PBL is increasing. Analysis of the learning process as well as appropriate support and organization of the experience into systematic knowledge and insight are key factors for achieving more effective learning. This research project generates a new program design and a new support system.

3 Requirements for the Collaboration Support Tools

Figure 1 shows the process of developing embedded software. Most existing training courses incorporating e-Learning technology focus on the right half of the process, which attempts to help learners acquire programming skills. Simulators, virtual hardware environments, and version control systems are often used for this purpose.

Without ignoring the right half of Fig. 1, the authors intensively focus on the left half of the development process and the postmortem process, which enables learners to share knowledge about designing embedded software and to strengthen insight for executing a project. For these purposes, ICT is required to handle and process the information and knowledge contributed by the learners. Therefore, knowledge technology for collaborative learning support targeting PBL of embedded software design has been the main focus of our research.

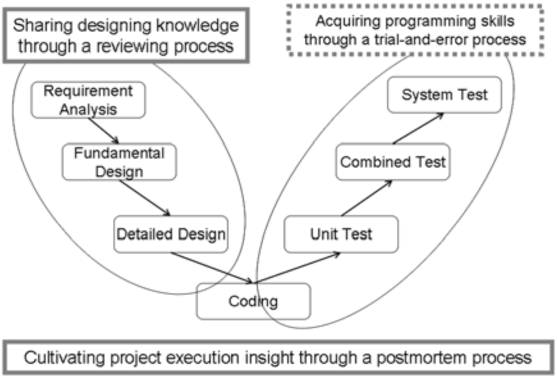

Sharing designing knowledge through a reviewing process
Acquiring programming skills through a trial-and-error process
Requirement Analysis
Fundamental Design
Detailed Design
Coding
Unit Test
Combined Test
System Test
Cultivating project execution insight through a postmortem process

Fig. 1. The Development Process of Embedded Software

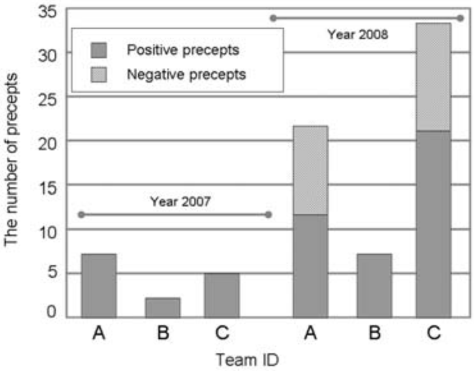


Fig. 2. Precepts Obtained Through the Postmortem Process in the PBL

The rest of this section describes the findings through observation of an actual PBL course. The findings clarify the requirements for the collaboration support tools. The learners' activities and communication were captured in the real PBL training course. The 10-day training course was held between October, 2008 and January, 2009. The 18 participants developed software for a remote controlled rover.

Documents and program codes created through the learning process were stored in a file server, and the modes of activities and communication between learners were recorded. By analyzing the captured data, the following findings for the collaborative learning support system were obtained.

1. The learners are trained by the experience of carrying out the project and applying their own knowledge and skills for software development and project management. Frequent verification and feedback would improve the learning effect. Cross review, in which a pair of learners review and annotate each otherf s design, and face-to-face group review are necessary in each step of the development process, although each learner conducts the design process

individually. Therefore, a function for performing cross and group reviews, even if the learners are in remote places, is required.

2. In the PBL course, the learners gain a sense of accomplishment with completion of the project; however, this fact itself does not contribute to the learners' ability. For achieving a learning effect, the experiences should be organized into structured knowledge. Therefore, the postmortem study is very important. Records of the learners' activities and communication are expected to make the postmortem more productive. To verify this assumption, in the postmortem phase, the participants were requested to write down DOs and DON'Ts (precepts) on the development of the embedded software, which they learned through PBL. The numbers of precepts are compared in Fig. 2. This research was conducted in 2007 and 2008. The learners in 2008 looked back at their learning process by using the records, whereas the learners in 2007 did not trace their development process. As shown in Fig. 2, the number of precepts for each team in 2008 is greater than that in 2007. Therefore, recording learning activities, retrieving them, and showing them in a time series would enable the learners to look back, discuss their design process, and organize the acquired knowledge.
3. In the face-to-face group discussion, use of a white board was very effective, and continuous presentation of the discussion process was rated as a positive effect for productivity and quality of the software. Therefore, an online white board that has a history function is required.

4 Online Collaboration Support Tools

Based on the above findings, an integrated repository tool and a unified search tool are proposed.

The integrated repository tool is the integration of file storage and a bulletin board. It is basically a BBS that is able to store multiple files as attachments of a posted message. Obviously, this tool can be used as a simple asynchronous communication tool; it can also be used for sharing documents and programs (products) created through the learning process and annotating the products. This would enable remote parties to perform cross and group reviews. The tool also has multiple display modes for the attached files, including an in-line display mode and a separate window display mode. Learners can choose the in-line display to obtain an overview of the design process and a separate window mode to investigate each of the product files. The repository tool is integrated in OpenSourceLMS [7]. A screen shot of this tool is shown in Fig. 3.

For the postmortem process, the relations between the products, learning contents, and discussion records should be found before looking back at the project history. The unified search tool can retrieve learning contents, stored documents, and messages posted at specific times. It is also integrated in OpenSourceLMS, as shown in Fig. 4.

Using these tools, project-based learners can frequently review their designs with each other, even if they cannot meet together in a physical classroom. They

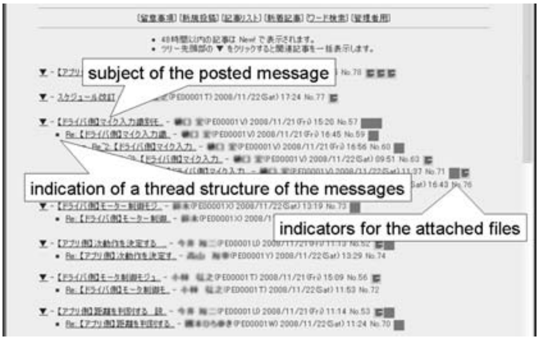

subject of the posted message
indication of a thread structure of the messages
indicators for the attached files

Fig. 3. Screen Shot of the Integrated Repository Tool

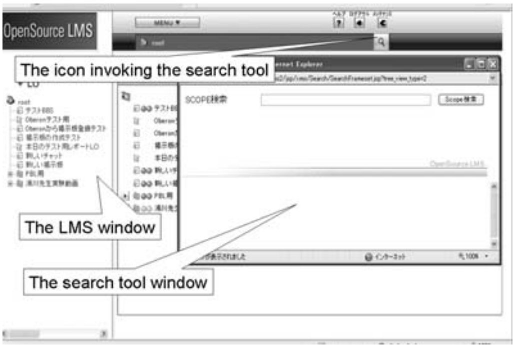


Fig. 4. Screen Shot of the Unified Search Tool

can also look back on their activities efficiently in the postmortem phase because the activities are recorded in the online repository and easily extracted with the unified search tool. Therefore, the tools are expected to improve the educational effect on project-based learners.

5 The Trial Training Course and Its Evaluation

The authors prepared a PBL training course as a trial. The target subject was development of an electronic thermometer using the one-chip microcomputer board, Renesas Starter Kit RSK M16C62P [8]. The 5-day course focused on the left half of the development process shown in Fig. 1: the requirement analysis, fundamental design, and detailed design. Due to the time limitation, the scope of the course excluded the coding and debugging phases. Nine students participated in the course and each student completed the design of the target thermometer.

In the course, the learners iterated the following steps:

1. Listen to the face-to-face lecture;
2. Carry out work individually and upload the design documents; The learner can look at other learners' progress and refer to other learners' documents;

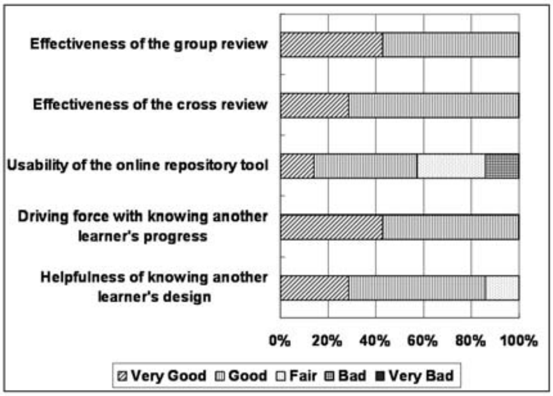


Fig. 5. Evaluation Results of the Trial Training Course

3. Perform the online cross review, in which a pair of learners review and annotate their each otherfs design by using the online repository tool;
4. Perform the face-to-face group review.

After completing the trial course, a questionnaire survey was conducted to evaluate the blended PBL course itself and the online repository tool. The questions asked about the effectiveness of the cross and group reviews, usability of the online repository tool, effectiveness of knowing another learner's progress, and helpfulness of knowing another learner's design. The results are shown in Fig. 5. As shown in the figure, every learner gave a positive response for the effectiveness of the reviews. This means that the online and face-to-face review processes are comparably effective. The online repository tool was easy to use for more than half of the learners, but some learners felt it was inconvenient. Improvement of usability of this tool is a future task. Knowing another learnerfs progress encourages the learner to work hard, and referring to another learnerfs design helps to improve the quality of the design. These results are achieved using the online repository tool; therefore, the blended learning program seems to have a beneficial learning effect.

6 Summary and Future Work

The present paper describes a research project to implement e-Learning technology and PBL online collaboration support tools for the development of embedded software. Through observation of a real PBL training course, the following functional requirements for collaboration support were clarified: a function for performing remote cross and group reviews, a function recording learning activities and displaying them in a time series, and an online white board with a history function. Based on these findings, an integrated online repository tool and a unified search tool are proposed.

The online repository tool is essentially a BBS that is able to store multiple files as attachments of a posted message. This tool can be used for sharing and annotating documents and programs. The usability and effectiveness of the online repository tool are evaluated through the trial PBL training course. The results of the questionnaire survey suggest that the tool is easy to use for most learners and is effective for conducting training collaboratively.

Implementation and integration of the white-board system with a history function, along with evaluation of the effectiveness of the unified search tool and white-board system, are areas for future study. The proposed tools support every step in the design phase presented in Fig. 1 because the essential characteristics of the steps are similar and necessary supports are almost identical. However, there are differences at a more detailed level. Therefore, additional functions dedicated for each of the steps would be effective, and these are also areas of future study.

References

1. Yukawa, T., Takahashi, H., Fukumura, Y., Yamazaki, M., Miyazaki, T., Yano, S., Takeuchi, A., Miura, H., Hasegawa, N.: Implementing e-learning technology for project-based learning for the development of embedded software. In: Proceedings of 20th annual conference of the Society for Information Technology and Teacher Education, pp. 2208–2212 (2009)
2. Niigata Industrial Creation Organization: Nico advanced training programs (2009), http://www.nico.or.jp/modules/list2/kouza_kensyu/koudoit.html (in Japanese)
3. Alavi, M.: Computer-mediated collaborative learning: An empirical evaluation. MIS Quarterly: Management Information Systems 18(2), 159–174 (1994)
4. Brandon, D.P., Hollingshead, A.B.: Collaborative learning and computer-supported groups. Communication Education 48(2), 109–126 (1999)
5. Gillet, D., Nguyen-Ngoc, A.-V., Rekik, Y.: Collaborative web-based experimentation in flexible engineering education. IEEE Transactions on Education 48(4), 696–704 (2005)
6. Kojiri, T., Kayama, M., Tamura, Y., Har, K., Itoh, Y.: Cscl and support technology. JSiSE Journal 23(4), 209–221 (2006) (in Japanese)
7. eLearning Consortium: Scorm2004 engine (2004), http://www.elc.or.jp/cgi-bin/scorm_engin/lms/index-scorm.html (in Japanese)
8. Renesas Technology Europe Ltd.: Renesas Starter Kit RSKM16C62P User's Manual (2007)

The Relationship between the Learning Styles of the Students and Their e-Learning Course Adaptability

Kazunori Nishino, Hiroko Toya, Shinji Mizuno, Kumiko Aoki,
and Yoshimi Fukumura

Kyushu Institute of Technology, Faculty of Computer Science and Systems Engineering,
680-4 Kawazu, Iizuka, Fukuoka, 820-8502 Japan
nishino@lai.kyutech.ac.jp, toya@smile.kyutech.ac.jp,
s_mizuno@aitech.ac.jp, kaoki@code.u-air.ac.jp,
fukumura@oberon.nagaokaut.ac.jp

Abstract. This study investigated learning styles of students who had or had not taken e-learning courses, developed a learning style questionnaire for e-learning courses, and examined the relationship between the learning style and the adaptability to e-learning courses. As the result, the student's adaptability of e-learning courses can be suggested before his/her taking an e-learning course. It was found that using the multiple regression model obtained in the study, about 40% of the adaptability to e-learning courses can be explained by the learning style questionnaire developed in the study.

Keywords: learning styles, e-learning, class adaptability, asynchronous learning.

1 Introduction

In recent years, e-learning has become widely used in higher education and company training. E-leaning can be classified into two categories based on the proportion of online components in a class: blended learning where e-learning supplements in-class teaching, and full online learning which is conducted without any face-to-face meeting. In this study, e-learning means the latter; full online learning.

E-learning has an advantage of allowing for learners to study at "any time" and "any place." In addition, it allows diverse learning forms with the use of information and communication technologies (ICT). On the other hand, as e-learning is usually conducted asynchronously, it requires more self-discipline of students in comparison to face-to-face classes. It might be easier for students who like to learn at their own pace to continue and complete e-learning. However, it can be challenging for those who don't like studying on their own and prefer face-to-face classes. It is not unusual that students drop out from e-learning courses and fail to get credits for the course they have registered for [1].

Several measures for preventing students' dropout have been taken with the use of ICT. The use of learning management systems (LMS) can ease the distribution of course materials and communication among students or between students and staff. Some measures have been taken to help students understand the content of e-learning

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 539–546, 2009.

materials and to motivate students in studying materials through e-mails sent by teachers and tutors of e-learning courses [2]. However, the use of ICT in e-learning tends to become complex as its functionality increases and may discourage those students who are not well equipped with the ICT use.

The use of ICT and asynchronous learning is a typical characteristic of e-learning. However, as it is stated earlier, those who don't like asynchronous learning or the use of ICT may have the tendency to drop out in the middle of e-learning courses. Therefore, it is desirable that students and their teachers know the students' learning styles and their adaptability of e-learning courses in advance [3].

This study aims to investigate the relationship between students' learning styles in e-learning and their adaptability of e-learning courses by surveying learning styles of students and analyze the correlation between the learning styles and the adaptability of e-learning courses.

2 Learning Styles in e-Learning

The research on learning styles has been popular in Western countries, especially in the U.K. and in the U.S. in the past 30 years. There has been a vast amount of literature on learning styles. According to the Learning Skills Research Center (LSRC) in the U.K., the number of journal articles on the subject has reached more than 3,800. In those articles, 71 different theories and models of learning styles have been presented. LSRC has selected 13 most prominent theories and models of learning styles among the 71 theories and models, and further studied the 13 models [4]. LSRC classified the 13 models of learning styles into five categories from the most susceptible to environments to the least susceptible ones based on the Curry's onion model [5].

As discussed above, there have been many theories about learning styles and it has not been agreed upon the flexibility of learning styles in terms of whether individual learning styles are independent of the learning environments or they are adaptable to the environments.

When investigating learning styles in e-learning, how should we consider the "flexibility of learning styles"? E-learning has the potential to provide "student-centered learning" and tends to be designed based on the pedagogy of providing learning environments according to the students' needs, abilities, preferences and styles rather than providing uniform education without any consideration of individual needs and differences. Therefore, it is meaningful to provide students and teachers with information about the students' adaptability of e-learning courses by using a questionnaire for learning styles in e-learning.

Though some studies were conducted on the Kolb's learning style [6] in developing computer-based training (CBT) [7], few studies on learning styles in e-learning have been done in the past. In this study, we developed a questionnaire to measure learning styles in e-learning, investigated the relationship between learning styles in e-learning and the adaptability to e-learning courses, and tested the validity of the questionnaire.

3 Development of the Questionnaire for Learning Styles in e-Learning

3.1 The Learning Style Questionnaire

To investigate the learning styles in e-learning, we developed a learning style questionnaire. The questionnaire consists of 36 items asking preferences in studying, understanding, questioning, and doing homework in terms of asynchronous learning and the use of ICT. In addition, the questionnaire included the four reverse coded items to test the validity of the data. The questionnaire was made available online as the target sample was the students who were taking e-learning courses and it was natural for them to access the questionnaire online.

The survey on learning styles was administered to those students who enrolled in the eHELP (e-Learning for Higher Education Linkage Project) which is a credit transfer system for e-learning courses offered by multiple universities in Japan. All the items in the questionnaire were asked with the 7-point Likert scale; from 1 being "don't agree at all" to 7 "agree strongly." The survey was conducted from the early December, 2008 to the early January, 2009, and obtained valid responses from 53 students. Those responses in which answers to the items were all the same including the reverse coded items were considered invalid.

3.2 Factor Analyses of the Questionnaire Results

A factor analysis of the data was conducted with SPSS using the Maximum Likelihood Estimation with Varimax rotation. The result is shown in the Table 1 below.

As the majority of the items which belong to the factor 1 concern the place, time, and content of asynchronous learning, the factor 1 is named as "preference for asynchronous learning." As for the factor 2, the majority of its items are about the use of computers in terms of studying and understanding, and this factor is named as "preference for the use of computers in learning." As the 9 items which belong to the factor 3 are mostly related to communication matters, the factor 3 is named as "preference for asynchronous digital communication." The items for the factor 4 concern the autonomy of deciding study sequence, and it is named as "study sequence autonomy."

The results of the factor analysis described above were analyzed in terms of the reliability of the factors. The four items which did not belong to any of the four factors mentioned above were excluded from the analysis. In addition, the factor 4 was also excluded from the analysis as the factor had only two items and did not contribute much to the overall explanation.

To test the reliability of each factor in the questionnaire, Cronbach α was analyzed for each factor. The Cronbach α for the factor 1, 2, and 3 resulted in 0.862, 0.805, and 0.664 respectively. As for the factor 3, if the item q34 was deleted the overall reliability would increase to 0.759. Therefore, the item q34 in the factor 3 was deleted in the further analysis.

As a result, the 33 items in the questionnaire comprise the three factors: the factor 1 "preference asynchronous learning," the factor 2 "preference for the use of computers in learning" and the factor 3 "preference for asynchronous digital communication," and those were analyzed further.

Table 1. The Result of Factor Analysis of e-Learning Learning Style Questionnaire Data (After Rotation)

	factor			
	1	2	3	4
q1) I understand better when I study at my convenient time rather than learning in class with other people. (async)	**0.891**	0.084	0.115	-0.099
q2) I can familiarize myself better when I study independently at my convenience than studying with others at one place. (async)	**0.729**	0.163	-0.202	-0.117
q3) I would rather study alone at the place and time convenient to me than learn in class with other people. (async)	**0.721**	-0.045	0.233	-0.041
q4) I can be more creative when I study alone than studying with others at one place. (async)	**0.694**	0.196	0.137	-0.085
q5) I feel more motivated when I study at my convenience than learning in class with other people. (async)	**0.681**	0.271	-0.009	0.132
q6) I can learn better when I study at the time I decide than when I study at the time decided by others. (async)	**0.681**	0.104	0.214	-0.086
q7) I tend to learn more actively when I study alone than studying with others at one place. (async)	**0.635**	0.028	0.008	0.052
q8) I study at my own pace and do not care how others study. (async)	**0.596**	-0.117	0.417	-0.142
q9) I can concentrate better when I study independently at my convenience than studying with others at one place. (async)	**0.594**	0.238	0.046	-0.11
q10) I feel less tired when I study independently at my convenience than studying with others at one place. (async)	**0.593**	-0.043	0.348	-0.078
q11) I wan to study at the same pace with other sudents. (sync)	**-0.587**	-0.031	0.032	0.153
q12) When I study through computers, I tend not to care how others study. (with ICT)	**0.564**	0	0.476	-0.075
q13) I wan to study at my own pace. (async)	**0.538**	0.285	0.106	0.262
q14) I tend to learn more actively using computers than studying in class. (with ICT)	0.358	**0.642**	0.075	-0.025
q15) It is easier for me to memorize what is on a computer rather than to review printed materials. (with ICT)	0.119	**0.635**	0.04	-0.247
q16) I can be more creative when I think on paper than using computers. (without ICT)	0.082	**-0.618**	-0.165	0.083
q17) I would rather do group learning through computers than face-to-face. (with ICT)	-0.012	**0.617**	-0.092	0.288
q18) I can concentrate better looking at a computer screen than looking at a blackboard or a large screen in a classroom. (with ICT)	0.452	**0.571**	0.069	0.113
q19) I feel more motivated when I study using computers than learning from teachers in person. (with ICT)	0.206	**0.563**	-0.157	0.15
q20) I understand better when I learn through computers than when I learn by reading books. (with ICT)	0.17	**0.562**	0.511	0.077
q21) I can be more creative when I think using computers than thinking on paper. (with ICT)	-0.018	**0.56**	0.277	0.01
q22) It is easier for me to communicate through computers or cell phones than to communicate face-to-face. (with ICT)	-0.087	**0.52**	0.39	-0.03
q23) I would rather follow the computer instruction rather than study reading textbooks. (with ICT)	0.208	**0.491**	0.458	-0.023
q24) I prefer learning through computers to learning by reading books. (with ICT)	0.221	**0.474**	0.447	0.102
q25) I feel less tired looking at a computer screen than looking at a blackboard or a large screen in a classroom. (with ICT)	0.076	**0.437**	0.245	0.034
q26) It is easier for me to take test on a computer than on paper. (with ICT)	-0.01	0.296	**0.599**	0.146
q27) I would rather submit my report in an electronic format than in a paper and pencil format. (with ICT)	0.083	-0.012	**0.588**	0.21
q28) It is easier for me to take test individually than to take one in a place with others. (async)	0.269	-0.006	**0.535**	-0.173
q29) I would rather receive answers later from teachers via mail than asking questions in person or through chat. (async)	0.001	0.161	**0.526**	-0.011
q30) I prefer communicating via email to communicating through telephones. (async)	-0.099	0.068	**0.468**	-0.086
q31) I am familiar with computers. (with ICT)	0.145	0.001	**0.461**	0.029
q32) I prefer taking notes using a computer than writing on paper. (with ICT)	0.204	0.336	**0.445**	-0.08
q33) I would rather ask questions using email or bulletin boards than asking teachers in person. (with ICT)	0.096	0.295	**0.445**	0.077
q34) I would rather study reading textbooks rather than follow the computer instruction. (without ICT)	-0.048	-0.67	**-0.422**	0.163
q35) I want to decide the study sequence on my own. (async)	0.27	-0.124	0.161	**-0.941**
q36) I want to follow the study sequence which my teacher decides. (sync)	-0.125	0.053	0.032	**0.583**
q37) I prefer being assessed individually upon completion of the assignment to being assessed at the same time with others. (async)	0.114	-0.22	0.091	0.238
q38) I want to drill what I have learnt repeatedly. (async)	0.259	-0.096	0.062	0.184
q39) It is easier for me to tackle with the project I decide than the one assigned to me. (async)	0.188	0.105	0.309	-0.194
q40) I prefer looking my grade online to being given it on paper. (with ICT)	0.212	0.343	0.372	-0.204

4 The Survey on the Adaptability to e-Learning Courses

When the learning style questionnaire was administered, the questionnaire on the adaptability to e-learning courses was also administered to the students who enrolled in eHELP courses. The items in the questionnaire are shown in the Table 2. The questionnaire consists of 10 items asking psychological aspects of learning such as the level of students' understanding and the level of satisfaction.

The questionnaire (see Table 2 below) was administered online to the students enrolled in each of the eHELP courses upon their completion of the course (i.e., between December, 2008 and January, 2009) and 69 responses completed the questionnaire. All the items in the questionnaire were asked with the 7-point Likert scale; from 1 being "don't agree at all" to 7 "agree strongly." The scores for the item (g) and (h) were reverse-coded. The mean score was 4.7.

To test the reliability of the 10 items in the questionnaire, Cronbach α was analyzed. As a result, Cronbach α = 0.783 was obtained and we determined to use all the 10 items as one factor of adaptability to e-learning courses.

Table 2. The Question Items in the Adaptability to e-Learning Course Questionnaire and Mean Scores

Item	Mean
(a) The content of this e-learning course is more understandable than regular class contents.	4.51
(b) The style of learning of this e-learning course is easier to learn than regular class.	4.90
(c) The pace of this e-learning course is more suitable than regular class.	4.91
(d) This e-learning course is more satisfying than regular class.	4.36
(e) This e-learning course is more effective than regular class.	4.35
(f) This e-learning course is more interesting than regular class.	4.91
(g) This e-learning course makes me more tired than regular class.	4.84
(h) This e-learning course makes me more nervous than regular class.	5.59
(i) This e-learning course brings me more endeavor than regular class.	4.07
(j) This e-learning course brings me more motivation than regular class.	4.41

5 Results and Discussions

5.1 Results of the Learning Style Questionnaire

Based on the learning style questionnaire analysis discussed in the section 3, the mean score for each factor was calculated. In addition, for the comparison purpose, the same questionnaire was administered in the early February, 2009, to those students who had not enrolled in eHELP courses. The mean scores of 53 eHELP students and 39 non-eHELP students for each of the three factors: the preference for asynchronous

Table 3. Factor Scores for eHELP Students and non-eHELP Students

	eHELP Students (n=53)	Non-eHELP Students (n=39)
Factor 1: preference for asynchronous learning	4.59	4.09
Factor 2: preference for the use of computers in learning	3.93	3.23
Factor 3: preference for asynchronous digital communication	4.35	3.55

learning, the preference for the use of computers in learning, and the preference for asynchronous digital communication, are shown in the Table 3.

In comparing the two means of eHELP students and non-eHELP students in terms of their scores of their learning styles, one-way ANOVA was analyzed for each of the three factors. The result showed that the learning styles were significantly different between the two groups (p<0.01) and the scores of all the factors are significantly higher among eHELP students than among non-eHELP students.

5.2 Correlations

Correlations between the scores of the three learning style factors and the score for the adaptability of e-learning courses were analyzed among the 69 respondents who completed both of the two questionnaires. The correlation r were shown in the Table 4 below.

Table 4. Correlations between the Adaptability to e-Leaning Courses and the Leaning Style Factors

	r	*P*	*n*
Adaptability - Factor 1	0.53	< 0.01	69
Adaptability - Factor 2	0.60	< 0.01	69
Adaptability - Factor 3	0.29	0.015	69

A statistically significant (p <0.01) correlation was seen between the learning style factor 1 (the preference for asynchronous learning) and the adaptability of e-learning courses and between the factor 2 (the preference for the use of computers in learning) and the adaptability. The correlation between the learning style factor 3 (the preference for asynchronous digital communication) and the adaptability to e-learning courses is not as high; however, the correlation is statistically significant at the level of p=.05.

5.3 Multiple Regression Analysis

In order to further investigate the relationships between the adaptability to e-learning courses and each of the three factors of learning styles, multiple regression analysis was conducted. The results are shown in the Table 5.

Table 5. The Result of Multiple Regression Analysis

Variable Name	Regression Coefficient	*P*
intercept	1.82	**< 0.001
Factor1 (preference for asynchronous learning)	0.23	**0.0054
Factor2 (preference for the use of computers in learning)	0.45	**0.0003
Factor3 (preference for asynchronous digital communication)	0.01	0.9383
Multiple *R*-square	0.43	** < 0.001
n	69	—

**significant at p=0.01. *significant at p=0.05.

As shown in the Table 5, the regression coefficients of the Factor 1 and Factor 2 are relatively high and the p values are less than 0.01. However, as for the Factor 3, the regression coefficient is low and its p value is also low. It can be suspected that the multicollinearity is high between the Factor 2 and Factor 3. Therefore, another multiple regression was conducted excluding the Factor 3. The analysis resulted in

$$\text{Adaptability to e-learning courses} = 1.835 + 0.231 \times \text{Factor 1} + 0.450 \times \text{Factor 2}$$

However, in order to apply this multiple regression model to the non-eHELP student group, we have to first examine if the eHELP student group is not different from the non-eHELP student group. In other words, we have to examine if the learning styles have changed in the course of taking the e-learning courses. If so, we cannot use the learning style factors as the inherent preferences of students and determining factors for e-learning adaptability for students who have never taken e-learning courses. Therefore, we compared the learning styles of two eHELP groups: one being a group of students who have been taking e-learning courses since Spring 2008 and the other being a group of students who have just started taking e-learning courses in Fall 2008.

The result of one-way ANOVA of variance shows that the difference between the two student groups is not statistically significant; therefore, we can conclude that the taking e-learning courses is not likely to affect students' learning styles. It has been considered that students who prefer asynchronous learning and the use of computers in learning opt to take e-learning courses. The student's adaptability of e-learning courses can be forecasted before his/her taking an e-learning course, using the multiple regression model obtained in the study.

6 Conclusion

This study investigated learning styles of students who had or had not taken e-learning courses, developed a learning style questionnaire for e-learning courses, and examined the relationship between the learning style and the adaptability to e-learning courses. As at present only about 40% of the adaptability to e-learning courses can be explained by the learning style questionnaire, the questionnaire needs to be refined further in the future. In addition, in order to further understand the relationship between learning style factors and the adaptability of e-learning courses, a future study may want to include those who have dropped out from the course as respondents. Furthermore, it would be insightful to see the relationship between the adaptability to e-learning courses and the actual student performances as well as the relationship between the adaptability to e-learning courses and the students' experiences of e-learning courses in the past.

References

1. Kogo, C., Nakai, A., Nojima, E.: Relationship between Procrastination Tendency and Student Dropouts in e-Learning Courses. In: Research report of JSET Conferences, vol. 2004(5), pp. 39–44 (2004)
2. Fuwa, Y., Ushiro, M., Kunimune, H., Niimura, M.: Efforts toward the Establishment of Quality Assurances for Adults Students of Distance Learning on e-Learning System -Practice and Evaluations of Support and Advice Activities-. Journal of Multimedia Aided Education Research 3(2), 13–23 (2007)
3. Nishino, K., Ohno, T., Mizuno, S., Aoki, K., Fukumura, Y.: A Study on learning Styles of Japanese e-learning learners. In: 11th International Conference on Humans and Computers, pp. 299–302 (2008)
4. Coffield, F., Moseley, D., Hall, E., Ecclestone, K.: Learning Styles and Pedagogy in Post-16 Learning. In: A Systematic and Critical Review, Learning and Skills Research Center, London (2004)
5. Curry, L.: An Organization of Learning Styles Theory and Constructs. In: ERIC Document 235185 (1983)
6. Kolb, D.A.: LSI Learning-Style Inventory. McBer & Company. Training Resources Group, Boston (1985)
7. Henke, H.: Learning Theory: Applying Kolb's Learning Style Inventory with Computer Based Training. In: A Project paper for A Course on Learning Theory (1996)

Effectiveness of Engineering Solution Case Document Search Based on TRIZ Contradiction Matrix Theory

Koji Yamada, Motoki Miura, Tessai Hayama, and Susumu Kunifuji

School of Knowledge Science, Japan Advanced Institute of Science and Technology,
1-1 Asahidai, Nomi,
Ishikawa 923-1292 Japan
{ykouji,miuramo,t-hayama,kuni}@jaist.ac.jp

Abstract. We propose a method to manage documents of engineering solution case based on TRIZ contradiction matrix theory. The document of engineering solution case involves know-how and techniques for solving mechanical issue. Usually the documents are written by engineers and practitioners, and managed by a company for sharing and inheriting among employees. However, an engineer who lacks literacy cannot find the previous case documents due to the inadequate keyword selections. To solve the query issue, we introduce TRIZ contradiction matrix theory for categorizing case documents. The engineers can retrieve adequate case documents by selecting improvement parameter and deterioration parameter on the matrix. Since the classification based on the matrix substantially categorize the case documents in terms of the problem solving methodology, it is effective and straightforward way of the specialized field and the key word. It is construction of the knowledge management support system that applies the idea of TRIZ. The problem solving that uses the reference information on this system is practiced and effectiveness is verified.

Keywords: TRIZ, Knowledge management support system, retrieval, contradiction matrix.

1 Introduction

Using the case collection made in the past case is important on knowledge and technological experience lore, and it has come to obtain a large amount of material easily by varied search engine in recent years. However, the literacy decrease's is due to decrease expert in specialized fields happening in enterprise etc. and obtaining a necessary case become difficult.

It is necessary to contain useful information on the success, the failure experience, and knowhow, etc., and to use the case collection for the power of people more. However, only the manufacturer understands the abounding knowledge, and it is likely not to transmit easily to the inexperience user of the case collection in the hoped real intention.

The conception knowledge of the word based on the expertise is necessary, and the case that the user requests is not necessarily obtained keywords by retrieval. Moreover, the number of case in each specialized field is limited, and similar solution of other fields cannot be used.

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 547–554, 2009.

In a general search process, the case is examined putting narrowing and retrieval by keyword according to the specialized field. However, the user has the doubt in utility that sees the case, and there is possibility of will not use case if it is the one that the keyword does not show the intention that the case originally has, too.

Then, this thesis examined the knowledge management support system by an approach different from the case retrieval existing. The knowledge management support system in a new aspect that applied the TRIZ theory to the case retrieval and the document management of the case collection was constructed so that the user might obtain a necessary case, and effectiveness was evaluated. The improvement can be expected by applying the idea of technological contradiction matrix to use an existing combination in the TRIZ theory and to propose solving, and giving the user the reference information with utility in respect of the quality of the problem solving in this system.

2 Proposed Method

The effort on the input method of the case title and the keyword registration type, etc. is done to use potential data of the case better so far, and it can be used though even the inexperienced person can easily retrieve. Even if there are a lot of technical terms, the interpretation is not a little difference. Consequently, related keyword included in the document is not exactly useful information.

How the content of the case is handled for the knowledge management support system that constructs it in the present study becomes the key. It is important to drop the superfluous information of the content of the case very, and to catch essence, and it is a problem how to link an actual case with the tool that can be the handling it.

Then, the present study examined the knowledge management support system by an approach different from the case retrieval existing. It wanted to apply the TRIZ

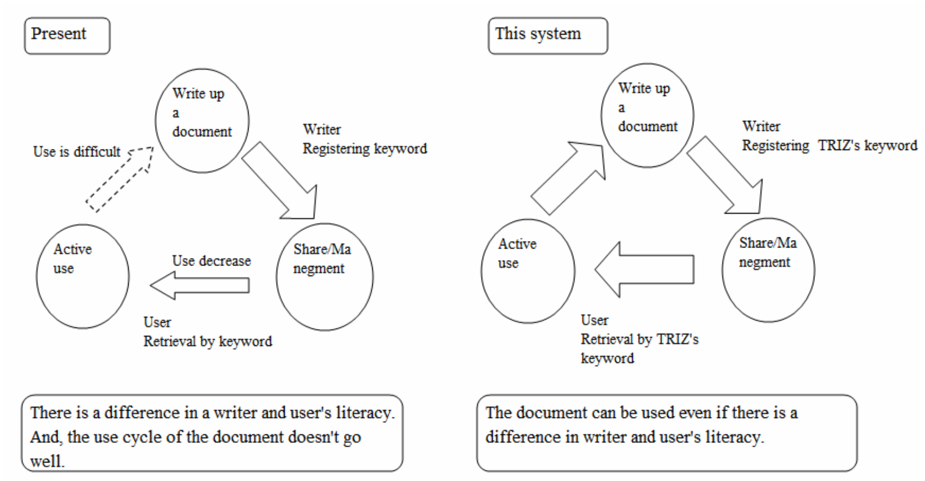

Present
Write up a document
Writer Registering keyword
Use is difficult
Active use
Use decrease
Share/Ma negment
User Retrieval by keyword
There is a difference in a writer and user's literacy. And, the use cycle of the document doesn't go well.
This system
Write up a document
Writer Registering TRIZ's keyword
Active use
Share/Ma negment
User Retrieval by TRIZ's keyword
The document can be used even if there is a difference in writer and user's literacy.

Fig. 1. Document use flow

theory to the case retrieval and the document management of the case collection so that the user may obtain a necessary case, to construct the knowledge management support system by a new viewpoint, and to evaluate effectiveness. Figure 1 is a conceptual diagram of the difference of the document use in current document use flow and the present study.

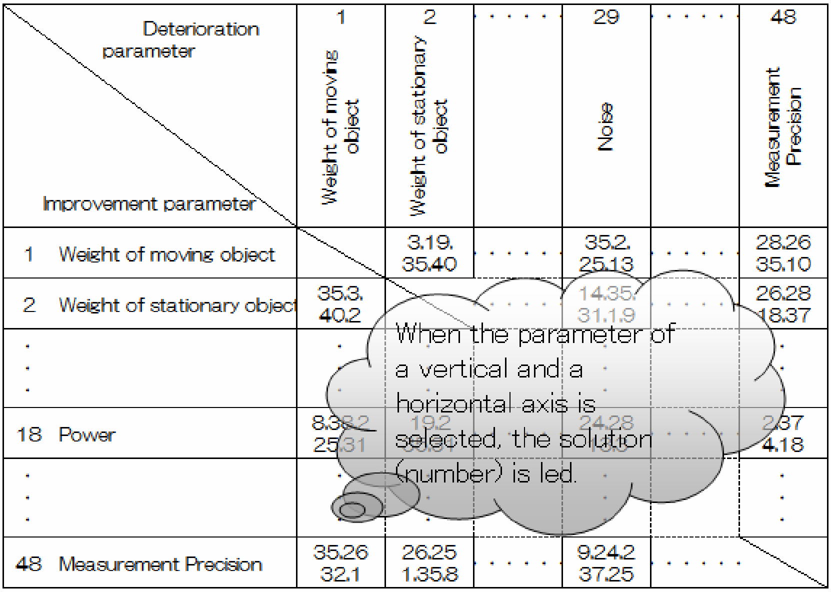
It is made easily to construct the one that the document use is promoted as being in the new flow of Figure 1 for the effectiveness of the problem solving to practice the document use in the present study and to verify it with the Web system of open source, to expand the user range, and to access information. It is applied to an actual case, and it is assumed that the quality of the problem solving is evaluated.

3 TRIZ Contradiction Matrix

It is time when actually occur a technical problem holds technical contradiction "Other one deteriorates by improving certain". At that time, the solution comes to do the trade-off (compromise), and to be going to grow dim, too. Because it persists in an immediate factor, it is thought that the directionality to the solution narrows and the flexibility of the idea is lost.

The TRIZ technology contradiction matrix is led to the solution by the parameter and catching, and replacing "Other one deteriorates by improving certain" with the item that abstracts the parameter respectively. Table 1 shows the outline of the TRIZ technology contradiction matrix.

Table 1. TRIZ Technology contradiction Matrix (Matrix2003)



Deterioration parameter Improvement parameter	1 Weight of moving object	2 Weight of stationary object	· · · · ·	29 Noise	· · · · ·	48 Measurement Precision
1 Weight of moving object		3.19. 35.40	· · · · ·	35.2. 25.13	· · · · ·	28.26 35.10
2 Weight of stationary object	35.3. 40.2		· · · · ·	14.35. 31.1.9	· · · · ·	26.28 18.37
· · ·	· · ·	· · ·		· · ·		· · ·
18 Power	8.38.2 25.31	19.2 [illegible]	· · · · ·	24.28 [illegible]	· · · · ·	2.37 4.18
· · ·	· · ·	· · ·		· · ·		· · ·
48 Measurement Precision	35.26 32.1	26.25 1.35.8	· · · · ·	9.24.2 37.25	· · · · ·	

When the parameter of a vertical and a horizontal axis is selected, the solution (number) is led.

The TRIZ technology contradiction matrix is classified into 48 items to which the parameter described in a spindle and a horizontal axis is common, and the solution led there is classified into 40 pieces, and the priority level is presented in the solution. A feature thing is to catch the true nature of the problem when replacing it with the

Table 2. Improvement and Deterioration Parameter (Matrix2003)

1	Weight of moving object	Physical
2	Weight of stationary object	
3	Length of moving object	
4	Length of stationary object	
5	Area of moving object	
6	Area of stationary object	
7	Volume of moving object	
8	Volume of stationary object	
9	Shape	
10	Amount of Substance	
11	Amount of Information	
12	Duration of action - moving object	Performance
13	Duration of action-stationary object	
14	Speed	
15	Force / Torque	
16	Use of Energy by moving object	
17	Use of Energy by stationary object	
18	Power	
19	Stress/Pressure	
20	Strength	
21	Stability	
22	Temperature	
23	Illumination Intensity	
24	Function Efficiency	Efficiency
25	Loss of Substance	
26	Loss of Time	
27	Loss of Energy	
28	Loss of Information	
29	Noise	
30	Harmful Emissions	
31	Object Generated Side Effects	
32	Adaptability / Versatility	Character
33	Compatibility / Connectability	
34	Ease of Operation	
35	Reliability	
36	Repairability	
37	Security	
38	Safety / Vulnerability	
39	Aesthetics	
40	Object affected harmful effects	
41	Manufacturability	Manufacturing and cost
42	Accuracy of manufacturing	
43	Automation	
44	Productivity	
45	System Complexity	
46	Device Complexity	
47	Ability to Detect/Measure	Measurement
48	Mesurement Precision	

Table 3. List of Invention Principle of TRIZ 40

1	Segmentation	21	Skipping
2	Taking out	22	Blessing in disguise
3	Local quality	23	Feedback
4	Asymmetry	24	Intermediary
5	Merging	25	Self-service
6	Universality	26	Copying
7	Nested doll	27	Cheap short-living objects
8	Anti-weight	28	Mechanics substitution
9	Preliminary anti-action	29	Pneumatics and hydraulics
10	Preliminary action	30	Flexibile shells and thin films
11	Beforehand cushioning	31	Porous materials
12	Equipotentiality	32	Color changes
13	The other way around	33	Homogeneity
14	Curvature	34	Discarding and recovering
15	Dynamization	35	Parameter changes
16	Partial or excessive actions	36	Phase transitions
17	Another dimension	37	Thermal expansion
18	Mechanical vibration	38	Strong oxidants
19	Periodic action	39	Inert atmosphere
20	Continuity of useful action	40	Composite materials

parameter of 48 items to obtain the requested solution for the problem when the TRIZ technology contradiction matrix is used. Therefore, it never enters when extra information retrieves it, and it can approach the requested solution.

Table 2 is a list of the parameter that composes a spindle and a horizontal axis of the TRIZ technology contradiction matrix.

As for 48 parameters, some classifications are performed. No.1-11 in "Physical", No.12-23, in "Performance", No.24-31, in "Efficiency", No.32-40, in "Character", No.41-46 in "Manufacturing and Cost", and No.47-48 are "Measurement".

Table 3 is the one that is called the invention principle of 40 that hits the solution of the TRIZ technology contradiction matrix. It is shown by the figure in the matrix in order of the use recommendation.

4 Outline of Knowledge Management Support System

Figure 2 shows the conceptual diagram of this system. This system replaces with the parameter that catches the true nature of the problem in the user, and inputs it to this system of PC. The content of the TRIZ technology contradiction matrix takes into SQL data base, comes to be able to display the solution on the Web system that makes it, chooses the solution that the user suits, and can extract the case in addition along it. What used the current having read it from the table of paper like Table 1 by hand power has been achieved by the automatic operation by Web. It is a composition of the output part where the solution is presented from the input part where the parameter of a vertical and a horizontal axis is selected on the Web menu and the input item and the retrieval part where the relating case data base is called there.

The user should work to the TRIZ technology contradiction matrix when replacing it with the parameter of abstract. The solution presented there is qualitative, and some experience and training are necessary for the replacement with the event of the real world from there. This system makes the case related to the parameter a data base to support the part, and it is possible to call it. The most this system particularly additional point is reference actually case by data base Web site. (For example: JST Failure Knowledge Database) In addition, the user can deepen the idea by referring to the presented case.

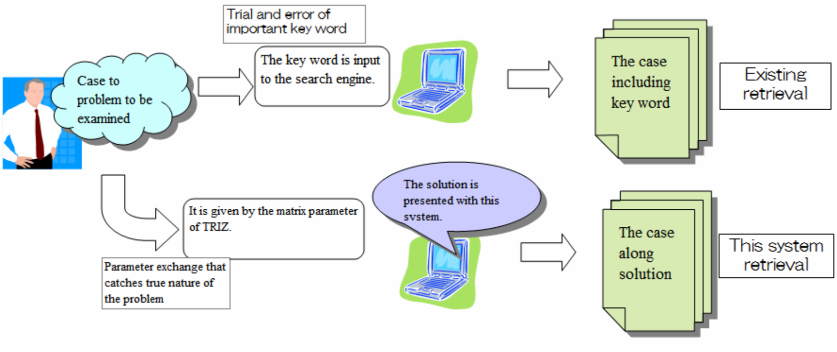
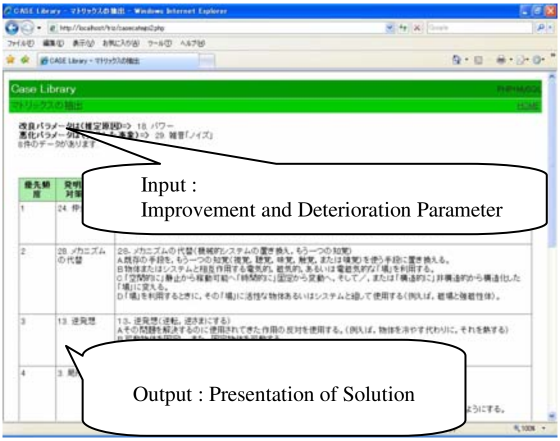

Trial and error of important key word
Case to problem to be examined
The key word is input to the search engine.
The case including key word
Existing retrieval
It is given by the matrix parameter of TRIZ.
Parameter exchange that catches true nature of the problem
The solution is presented with this system.
The case along solution
This system retrieval

Fig. 2. Existing retrieval and comparison of concepts of this system

5 Usage Scenario of the System

The solution is easily presented by selecting inputting the vertical and the horizontal axis parameter of the TRIZ technology contradiction matrix. The solution is put the order of priority, and decides which solution has suited the settlement of the issue by the user according to it, and can retrieve the case corresponding to it. For instance, it is assumed that the noise problem occurs though I want to raise the output of the amplifier. When the parameter that relates when the user inputs it to the system is selected, "Power" and the deterioration parameter become "Noise" as for the improvement parameter. After selecting the parameters, the system presents some solutions. The user selects the appropriate one from among that. It is hit on that the meaning "Intermediary" indicates digital processing and the noise removal functions etc. when thinking that the solution "Intermediary" is the most appropriate here. When the user doesn't hit on, knowledge where the system calls the case collection that relates to "Intermediary" can be supported. The solution of priority ranking is cause by TRIZ. Figure 3 is an example of the screen of presenting the solution after inputting the parameters.


CASE Library - マトリックスの抽出 - Windows Internet Explorer
http://localhost/triz/casecatego2.php
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
CASE Library - マトリックスの抽出
Case Library
マトリックスの抽出
HOME
Input :
Improvement and Deterioration Parameter
Output : Presentation of Solution

改良パラメータは(推定原因)=> 18. パワー
悪化パラメータは(…)=> 29. 雑音「ノイズ」
8件のデータがあります

優先順位	発明対策	
1	24. 仲…	
2	28. メカニズムの代替	28. メカニズムの代替(機械的システムの置き換え、もう一つの知覚) A既存の手段を、もう一つの知覚(視覚、聴覚、味覚、触覚、または嗅覚)を使う手段に置き換える。 B物体またはシステムと相互作用する電気的、磁気的、あるいは電磁気的な「場」を利用する。 C「空間的に」静止から稼動可能へ「時間的に」固定から変動へ、そして／、または「構造的に」非構造的から構造化した「場」に変える。 D「場」を利用するときに、その「場」に活性な物体あるいはシステムと組して使用する(例えば、磁場と強磁性体)。
3	13. 逆発想	13. 逆発想(逆転、逆さまにする) Aその問題を解決するのに使用されてきた作用の反対を使用する。(例えば、物体を冷やす代わりに、それを熱する)
4	3. 局…	…ようにする。

Fig. 3. Example of Displaying Solution of This System

6 Experiment and Evaluation

Five graduate students and seven company technical engineers participated in the experiment. Graduate students were inexperience persons, and technical engineers were experienced. The experience persons have knowledge of rotating machine techniques (drawing and calculate experience: fan, blower and compressor etc,). The technological trouble case that was actually was quoted from Nakao's failure best 100 and the experiment was prepared by two titles. Their title are “Failure of Return to the

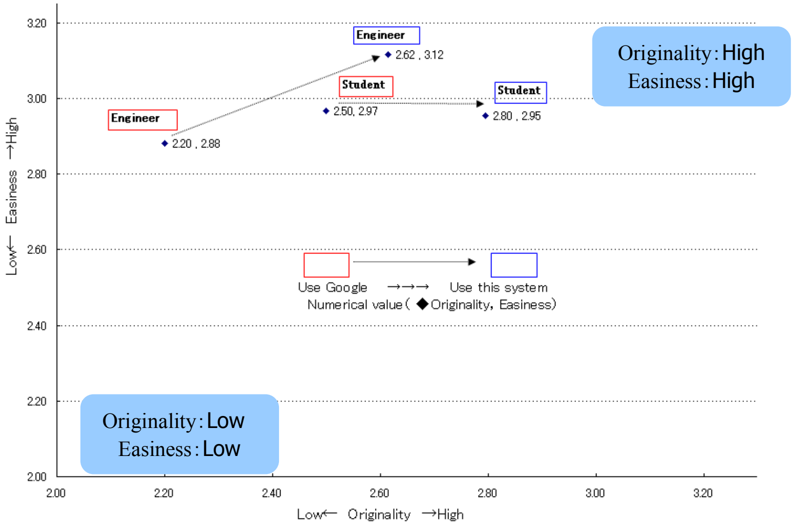


Fig. 4. Solution key word Evaluation of Originality - Easiness

Earth of Space Shuttle 'COLUMBIA'" and "The rust proof painting is defective in the coil spring". It searched for the solution by the index site by the first problem, and the solution was requested in the following problem according to the procedure of the conception of the system and TRIZ made in the present study. To prevent it with the rose of data by the difficulty of the problem, order was changed and alternately (1st - 2nd title or 2nd -1st title) executed by participants. And 1st system is used search engine, 2nd system is used this system. The evaluation score (five stages) was applied to each solution key word by originality and easiness as an evaluation of the utility of the solution key word, each mean value was compared.

Originality of solution key word as for easiness, the student was a level-off and a tendency that the evaluation raises in originality in an easy evaluation. It became a tendency that this system of both originality and easily rises in the Engineer. (Fig.4)

7 Conclusion

In this paper, a method to manage engineering documents by categorizing TRIZ contradiction matrix is presented. The engineers who cannot find proper query keywords can retrieve a set of document by selecting the parameters regarding the problem. We developed a system which implements the proposed method as Web service. We also confirmed the effectiveness of the proposed method through an experiment with both experts and non-experts of mechanics.

The merit of our method is to free the engineers from considering advanced technical terms while retrieving case documents. Thus this approach is suitable for non-expert engineers. Also referring the similar case documents in the same matrix cell inspires the non-expert engineers to find better solution and related terms. We continue to confirm the effectiveness of this approach by increasing case documents, and including documents of other areas.

References

1. Mann, D., Nakagawa, T.: Knowledge Creation Study Group: TRIZ Practices and Benefits. Systematic Technological Innovation, vol. 1 (2004) (in Japanese)
2. Mann, D., Nakagawa, T.: TRIZ Practices and Benefits. Vol.2. New Contradiction Matrix (Matrix 2003) (for Technologies in General) (2005) (in Japanese)
3. Nakao, M.: 100 Scenarios of Failure (2005) (in Japanese)
4. JST Failure Knowledge Database, http://shippai.jst.go.jp/fkd/Search

A Following Method of Annotations on Updated Contents and Its Evaluation

Hisayoshi Kunimune[1], Kenzou Yokoyama[2], Takeshi Takizawa[2], and Yasushi Fuwa[2]

[1] Faculty of Engineering, Shinshu University
[2] Graduate School of Science and Technology, Shinshu University
4-17-1 Wakasato, Nagano City, Nagano 380–8553 Japan
kunimune@cs.shinshu-u.ac.jp

Abstract. We have already developed an annotation sharing system for web-based learning materials. This system allows learners to write annotations such as markers and memorandums directly on materials and to share these annotations with lecturers and learners. If necessary, web-based materials can often be updated by authors; however, the annotations on the material are not in the proper position after the update. In the present study, we propose a method to follow the proper position of annotations in updated materials, and conduct experiments to evaluate the method. The following paper describes the proposed method and the experimental evaluation.

Keywords: annotation, e-learning, updated contents, following method.

1 Introduction

Web-based training courses are in widespread use by educational facilities and companies. These courses provide materials as web pages, and these web-based materials have the following useful features compared to paper-based materials:

- Multimedia contents such as movies, audios, and images can be included.
- The contents can be revised by lecturers at anytime.

Despite the popularity of web-based courses, learners still like to write annotation texts and marks on paper-based materials to aid their understanding[6]. However, if learners print these web-based materials to write annotations, they lose many of the benefits of web-based materials as mentioned above. That is, the printed materials do not include any multimedia contents and become out of date, whereas the web-based materials can be updated and revised at anytime. Many learners studying self-paced distance e-learning courses have problems related to printed web-based materials[3].

To solve these problems, several systems that allow the user to incorporate annotations in the web-based contents and to share these annotations have been proposed[1,2]. Learners can write annotations directly on web-based materials

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 555–563, 2009.

with these systems and browse the materials with multimedia contents. However, lecturers update the materials, and annotated sentences or words are moved and not displayed in the proper positions.

We proposed an annotation sharing system called *Writable Web* [3,4], which is implemented as a web server application. This system allows users to write annotations —marks on text, memos on text and images, and freehand drawings on images— on web-based materials and share them with lecturers and learners. Moreover, we propose a method to follow the proper positions of annotations, after the materials have been updated, using natural language processing techniques.

The present study examines a method to follow proper positions of annotations on updated materials, and describes the evaluation of the method according to experimental results.

2 Overview of Writable Web

Writable Web has the following features:

- Works on common web browsers and does not require the installation of any special software to run.
- Offers writing marks and memos on texts (Fig. 1) and freehand drawings and memos on figures (Fig. 2).
- Annotations are in proper positions after materials are updated.
- Supports online asynchronous discussion with shared annotations.

The architecture of the *Writable Web* system is illustrated in Fig. 3. The system works as a server-side web application between the web browser of a user and web servers that provide web-based learning materials. *Writable Web* works on commonly used web browsers, and the annotations that are written by each user are stored in the same database on the server, allowing users to share their annotations.

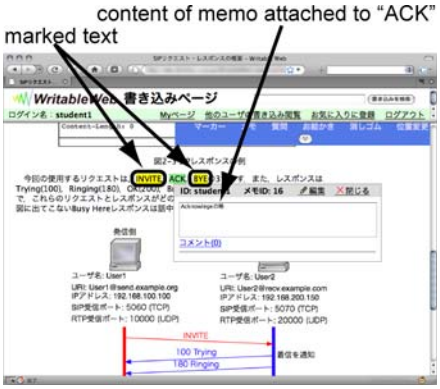


Fig. 1. Marks and memo on text

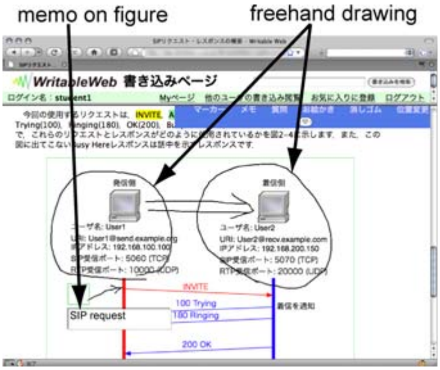


Fig. 2. Freehand drawings and memo on figure

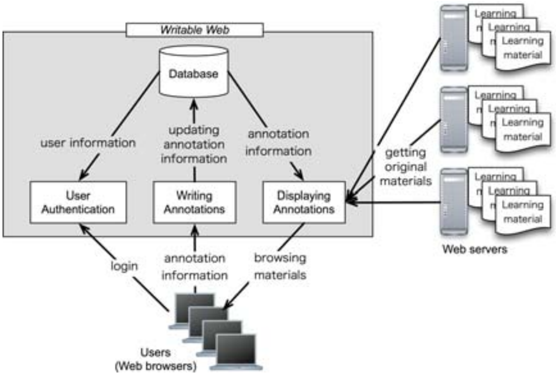


Fig. 3. Architecture of *Writable Web*

When a user writes an annotation on a page, *Writable Web* stores information which specifies the position, the type, the owner and the content of the annotation in a database. When users browse that page, the system inserts HTML tags to draw annotations into the original material only on memory, and does not make any changes to the original material itself. Moreover, the system stores the HTML file of the original material as a cache, and uses the cache, instead of the original material, when a user browses the same material, and will use the cache to detect for any updates.

3 Following Method of Annotation Positions on Updated Contents

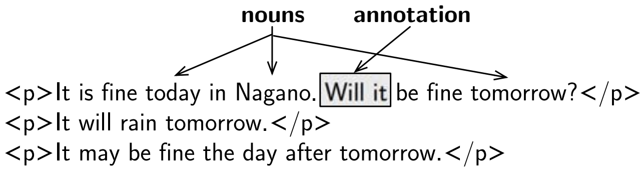
Lecturers sometimes update contents to refine their materials, and annotations written before updating are unintentionally misplaced. Then, some paragraphs, sentences or letters are added, deleted or moved by the lecturer, and the positions of annotated texts or words are changed. *Writable Web* stores only the physical position of annotations, which consists of the start and the end of annotations, such as the paragraph number on a web page and the letter number in a paragraph.

In the present study, we propose a method to estimate the proper position of paragraphs and letters that have moved following the update of web content. *Writable Web* detects the updates and automatically revises stored positions of annotations to the proper positions using this method.

The proposed method estimates the proper positions of annotations using information within the text as mentioned below.

3.1 Information to Estimate Proper Position

Information to estimate the proper position of an annotation consists of following elements in addition to the physical position of the annotation (L). N_b and N_a will be empty if there is no noun before and after an annotation in a paragraph.



<p>It is fine today in Nagano. Will it be fine tomorrow?</p>
<p>It will rain tomorrow.</p>
<p>It may be fine the day after tomorrow.</p>

Fig. 4. Example of information from annotation

$$\begin{cases} S & : \text{an annotated string} \\ N_b & : \text{the sequence of nouns before the annotation in a paragraph} \\ N_a & : \text{the sequence of nouns after the annotation in a paragraph.} \end{cases}$$

An example of the information from an annotation is shown in Fig. 4. In this example, each element of information is as follows:

$$\begin{cases} L & = \{\{1,29\},\{1,35\}\} \\ & (\{\{\text{start paragraph, start letter}\}, \{\text{end paragraph, end letter}\}\}) \\ S & = \text{"Will it"} \\ N_b & = \{\text{"today","Nagano"}\} \\ N_a & = \{\text{"tomorrow"}\}. \end{cases}$$

3.2 Estimating Proper Paragraph

The method estimates the proper paragraph in updated material as the first step in estimating the proper position of an annotation. Let $p_1, p_2, \cdots, p_l$ denote the string of each paragraph, where the updated material has l paragraphs. Let P' denote the paragraph most similar to the annotated paragraph before the update (P) as the proper paragraph. If the degree of similarity is lower than threshold, the method discontinues estimating the proper position.

The degree of similarity between paragraphs used in this method is proposed by Odaka et al. [5], as shown in formula (1), and shows the frequency of appearance of n-grams in two documents from zero (low similarity) to one (high similarity), with n as three in this method. The degree indicates high similarities against changing words, the ending of sentences, and the order of phrases; and Odaka et al. use it to measure the similarities between reports by students.

$$R_j = 1 - \frac{1}{k} \sum_{i=1}^{k} \left\{ \frac{P(X_i) - p_j(X_i)}{P(X_i) + p_j(X_i)} \right\}^2 ,$$

where
R_j : the degree of similarity of P and $p_j (1 \leq j \leq l)$ (1)
k : the total number of n-grams in P and p_j
X : the sequence of n-grams
$P(X_i)$: the frequency of appearance of X_i in P
$p_j(X_i)$: the frequency of appearance of X_i in p_j.

3.3 Estimating Proper Position in a Paragraph

This method estimates proper position of an annotation in the estimated paragraph (P') as follows:

Step 1. Find the same strings as annotated string (S) from P'. Let $s_1, s_2, \cdots, s_m$ denote the same strings as S in P'.

Step 2. Compare nouns before and after each found string ($s_1, s_2, \cdots, s_m$) with nouns before and after annotated string (N_b and N_a).
Let m_{bj} denote the number of nouns before s_j in the paragraph which matches nouns in N_b, and let m_{aj} denote the number of nouns after s_j in the paragraph which matches nouns in N_a. Let N denote the greater number of nouns before and after S or s_j in the paragraph.
Then, the degree of match (E_j) is defined as shown in formula (2) and has a value between 0 and 1.

Step 3. Decide the position of the string that has the greatest degree of match, as the estimated proper position (L').

$$E_j = 1 - \frac{N - (m_{bj} + m_{aj})}{N}, (1 \leq j \leq m). \tag{2}$$

If there is no string identical to S in P' in Step 1, or the value of the greatest degree of match is lower than threshold, there is a possibility that the annotated string has changed. As a result, the position of a substring, which has the greatest degree of match and the greatest degree of similarity, in the estimated paragraph (P'), is selected as the estimated proper position (L'). However, this method discontinues estimating the proper position when the degree of match or the degree of similarity is lower than threshold.

3.4 Thresholds

This method discontinues estimating the proper position depending on the degree of match or the degree of similarity, as mentioned in sections 3.2 and 3.3. At this point, users of *Writable Web* should manually move annotations to the proper position, and the system offers some functions to support this procedure. We think it is better to stop estimations rather than report incorrect estimations, so the system detects and notifies users of annotations that should be moved manually.

Using a low threshold, this method halts the estimation less frequently, but it incorrectly reports the estimation of annotation positions more frequently.

4 Evaluation

We conducted two experiments to evaluate the method to estimate the proper positions of annotations. The first experiment was conducted with annotations written by learners, and the second experiment was conducted with annotations which were randomly generated according to a trend in the annotations of learners. We confirmed the adequacy of the trend from the annotations of learners prior to beginning the second experiment.

Table 1. Result of following updates with annotations from learners

Estimation result	Number of annotations	Percentage
Proper	358	95.7%
Discontinued	13	3.4%
Incorrect	3	0.9%

Table 2. Result of exploratory experiment

Estimation result	Number of annotations	Percentage
Proper	381	93.2%
Discontinued	25	6.1%
Incorrect	3	0.7%

Table 3. Result of following updates with generated annotations

Estimation result	Number of annotations	Percentage
Proper	1285	61.0%
Discontinued	789	37.5%
Incorrect	32	1.5%

4.1 Following Updates with Annotations from Learners

This experiment was conducted to confirm the precision of the method with annotations written by learners, six undergraduate students and 19 master's students. They learned and wrote 1008 annotations on 11 pages of material using *Writable Web*. The material had been updated, including 14 revisions in one paragraph and the addition of five paragraphs; however, the learners were instructed to use the material prior to the updates, and were unaware of any changes in the material.

The results of this experiment are shown in Table 1. There were 374 annotations, 37.1% of all annotations, which were needed to follow updates. The types of estimation results are estimating the proper position, discontinuing the estimation, and estimating an incorrect position.

These results indicate that the precision of estimation using this method is very high. The high precision may be the result of relatively few updates in the material and updates that did not have an effect on other parts of the page.

4.2 Following Updates with Randomly Generated Annotations

We conducted the second experiment to increase the samples of updates. To increase the number of annotations, we planned to randomly generate annotations according to a trend in the annotations of learners, the morpheme number distribution. The distribution of 1008 annotations collected in the first experiment is shown in Fig. 5.

Exploratory Experiment. We conducted an exploratory experiment to confirm the adequacy of using this trend to generate annotations. Using the material from the first experiment, 1008 annotations were generated, starting from randomized morphemes, and the lengths of these annotations were decided according to the distribution.

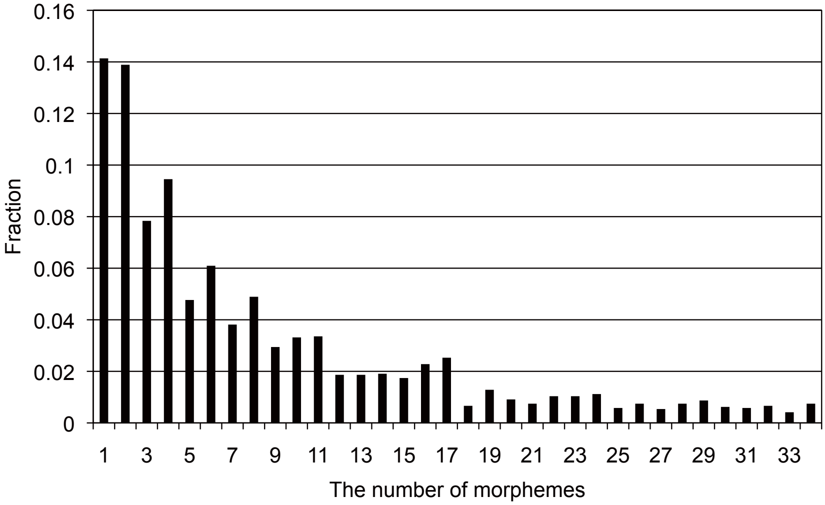


Fig. 5. Morpheme number distribution of learners' annotations

The results of the exploratory experiment are shown in Table 2. There were 409 annotations, 40.6% of all annotations, which were needed to follow updates.

There was a small difference in the result of discontinuing the estimation. The reason for the difference was that the number of annotations straddling multiple paragraphs was larger than in annotations made by learners, with some paragraphs being added between paragraphs, and the string in the annotation position changing. There were 198 annotations which straddled multiple paragraphs in the generated annotations, and there were 51 in annotations made by learners. However, the difference between the result of the first experiment and the exploratory experiment was small, so we used the generated annotations in place of the annotations from learners.

Experiment. We collected 30 materials before and after the update, and 3000 annotations were generated on pages before the update (100 annotations for each page). The results of this experiment are shown in Table 3. There was 2106 annotations, 70.2% of all annotations, which were needed to follow updates.

These results indicate that the number of annotations on which estimation was stopped was larger than in both the first and the exploratory experiments; however, the number of annotations that were misplaced was very low.

This experiment was conducted with a large number of annotations, which were randomly generated according to a trend from annotations made by actual learners; although, the trend used in this experiment was simple, and there was a difference in properties between annotations from learners and the generated annotations, as mentioned in the section 4.2. Confirming the precision of the estimation method is also very important when providing this method to actual learners.

Moreover, *Writable Web* offers functions to support manually moving the annotations to the proper positions, so we believe that there are no practical issues using this method.

During this experiment, we also checked the time to estimate the proper position of 100 annotations on each page. The mean and standard deviation of processing time were 119.2 seconds and 102.0, and the shortest and the longest processing times were 18 seconds and 557 seconds, respectively. On the page that took the proposed method 557 seconds (the longest processing time) to estimate, there were long annotations with more than 10 morphemes.

We believe the proposed method is very simple, and that we can improve the proposed method to estimate the proper position more quickly.

5 Conclusions

The present study proposed a following method of annotations on updated contents and described the result of experiments to evaluate the method. We found that the method can estimate the proper positions of annotations on contents having small updates, and confirmed that the method avoids moving annotations to incorrect positions by discontinuing the estimation.

We plan on improving the method by increasing the precision, decreasing the processing time of the method, and by tuning the thresholds. Moreover, we would like to evaluate the precision of the improved method by providing it to actual learners.

Acknowledgement

This research was partially supported by the Ministry of Education, Culture, Sports, Science and Technology, Grant-in-Aid for Young Scientists (B), 18700642, 2006–2008 and 21700808, 2009–2011.

References

1. Cadis, J., Gupta, A., Gruding, J.: Using Web Annotations for Asynchronous Collaboration Around Documents. In: Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work, pp. 309–318 (2000)
2. Koivunen, M.: Annotea and Semantic Web Supported Collaboration. In: Proceeding of UserSWeb, 1st Workshop on End User Aspects of the Semantic Web, vol. 137(1), pp. 5–17 (2005)
3. Kunimune, H., Yokoyama, K., Niimura, M., Fuwa, Y.: An Annotation Sharing System for e-Learning Materials. In: Proceedings of ED-MEDIA 2007, pp. 3366–3371. Association for the Advancement of Computing in Education, Norfolk (2007)

4. Kunimune, H., Yokoyama, K., Takizawa, T., Hiramatsu, T., Fuwa, Y.: A Web-based Asynchronous Discussion System and Its Evaluation. In: Lovrek, I., Howlett, R.J., Jain, L.C. (eds.) KES 2008, Part III. LNCS, vol. 5179, pp. 507–514. Springer, Heidelberg (2008)
5. Odaka, T., Murata, T., Gao, J., Suwa, I., Shirai, H., Takahashi, I., Kuroiwa, J., Ogura, H.: A Proposal on Student Report Scoring System Using *N*-gram Text Analysis Method. The IEICE Transaction on Information and Systems J86-D-I(9), pt. 1, 702–705 (2003)
6. O'Hara, K., Sellen, A.: A comparison of Reading Paper and On-line Documents. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 335–342 (1997)

Organization of Solution Knowledge Graph from Collaborative Learning Records

Yuki Watanabe, Tomoko Kojiri, and Toyohide Watanabe

Graduate School of Information Science, Nagoya University
Furo-cho, Chikusa-ku, Nagoya, 464-8603, Japan
{ywatanabe,kojiri,watanabe}@watanabe.ss.is.nagoya-u.ac.jp

Abstract. In collaborative learning, participants generate their own answers by exchanging their opinions through a discussion. Since the discussion in a collaborative learning includes knowledge for solving an exercise, the collaborative learning record is useful for other learners who tackle the same exercise. We propose a method for organizing solution knowledge in collaborative learning records as a solution knowledge graph. In this method, utterance collections of the same answering method are extracted and structured from a viewpoint of their effectiveness based on annotations attached by participants. In addition, the structure of the solution knowledge graph is refined by learning records of self-learners who use it as knowledge for solving exercises.

Keywords: solution knowledge graph, learning record, collaborative learning, self-learning, knowledge extraction, effectiveness of utterances.

1 Introduction

Recently, to support a collaborative learning under a distributed environment is one of the hottest subjects[1][2]. In the collaborative learning, participants propose their own ideas, ask questions, and reply to the questions in order to solve a common exercise. Since such utterances contain hints to solve the exercise, participants acquire knowledge to derive answers from the utterances which they cannot think of by themselves. Such knowledge is useful for other learners who did not participate into the collaborative learning, but tackle the same exercise.

Participants in different learning groups may derive answers with different answering methods. Even if they follow the same answering methods, they discuss differently with different knowledge. Therefore, it is useful for other learners to observe plural discussion records of the same exercise. However, if there are many discussion records for the same exercise, it is difficult for learners to find utterances that are appropriate for their learning situations. Our objective is to extract knowledge for deriving the same exercise automatically from the discussion records in the several collaborative learnings in order for other learners to utilize the discussion records as knowledge to solve the same exercise.

Several researches have been investigated for utilizing collaborative learning records. Kayama, et al.[3] developed a system which supports learners to review

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 564–571, 2009.

the contents of the collaborative learning according to tags that are automatically attached to specific actions of participants by the system. Goodman, et al.[4] constructed SAILE which organizes collaborative learning records based on status in common workspace. This system provides asynchronous collaborative learning environment where learners can review all actions, and replay to actions of the scenes that are defined by chat events. In addition, learners can increase the branch of a learning process by attaching comments to the discussion record. These researches allow learners to review collaborative learning record from specific scene. However, it is difficult to find scenes that may help them to derive the answer.

In order to extract parts which can be used for deriving answers from collaborative learning, it is necessary to detect topics. Adams, et al.[5] developed a system which extracts utterances of the same topic. This system calculates similarities between utterances according to feature vectors of utterances and detects topics. The feature vectors are estimated with frequency of terms in all utterances based on *tf-idf*. Okazaki, et al.[6] proposed a method for extracting and organizing topics from text documents. In this method, similarities between sentences are calculated with vocabularies, and sentences are organized according to their similarities. However, these researches only arrange topics in terms of similarities. Utterances in the collaborative learning should be organized from a viewpoint of their effectivenesses in solving exercise.

Currently, we focus on *self-learners* who did not participate into the collaborative learning and study individually using the collaborative learning records as hints for solving the exercise. Self-learners refer to the collaborative learning records when they cannot derive answers by themselves. Thus, utterance collections that are effective for solving the exercise need to be distinguished from other utterances of the same topic. Kakehi, et al.[7] developed a system which extracts useful utterance collections for deriving answers from a discussion record and organizes them along the time sequence. The system detects effective utterances based on annotations that are attached by participants to useful utterances for deriving their answers during the collaborative learning. This system could extract utterances that may be useful for solving the exercise, but could not show effectiveness of extracted utterances. Moreover, it is able to handle only a discussion record of a single collaborative learning.

In this research, effective utterances in the plural discussion records are extracted and structured as one solution knowledge graph according to the effectiveness for solving an exercise. Effective utterances may be referred by many participants/self-learners. In our approach, useful utterances are detected based on the participants' intentions for solving the exercise. As same as Kakehi's method[7], participants' intentions are grasped by annotations that are attached to utterances which are referred in solving the exercise. Effectiveness of utterances is inferred by the number of participants who attached annotations. Thus, extracted utterances are arranged in the solution knowledge graph according to the number of annotations.

The number of participants in the group may be small, so utterances that are annotated by participants are not always effective for other self-learners. Moreover, meanings of utterances in the discussion may not be the same for other self-learners, since they do not know the context of utterances. Therefore, the structure of the solution knowledge graph is refined based on the self-learners' learning records. By considering intentions of self-learners, for the solution knowledge graph, it is able to organize knowledge discussed in the collaborative learnings according to the importance in solving the exercise.

2 Approach

2.1 Collaborative Learning Environment

We focus on a collaborative learning of programming exercises. The solution of a programming exercise is composed of several answering steps which correspond to sub-exercises. For example, in an exercise of "construct a program which retrieves an input string from a file", there are there answering steps such as "to obtain input string", "to operate a file", and "to retrieve string from the file". Most sub-exercises are independent to others. Several answering methods can be applied to solve the sub-exercises. Moreover, each answering step holds keywords that can identify the step.

We assume the group of participants who have the similar understanding levels and try to solve the same exercise. Participants compose their own programs while discussing their ideas with others through a chat tool. When generating utterances, participants need to indicate target utterances. Also, they are required to attach annotations to utterances that are used for deriving their answers.

2.2 Framework for Organizing Knowledge in Discussion Records

When solving an exercise, self-learners cope with answering steps individually. Therefore, utterances of the same topic need to be extracted as an utterance collection and be corresponded to the discussed answering steps. In addition, in order for self-learners to find the effective hints quickly, utterance collections should be evaluated and ranked according to their effectiveness.

In a discussion record, several utterances may be generated for one topic. It is necessary to detect successive utterances of the same topic as an utterance collection and specify their answering steps. At this time, utterance collections are extracted from all the collaborative learning records for the same exercise. If more than one answering methods exist, differences among the answering methods should be specified. In addition, the effectiveness of utterance collections for deriving an answer may help learners to find useful hints for each answering method.

In order to organize solution knowledge from collaborative learning records, we design the learning environment which consists of mechanisms for extracting and structurizing the hints in the discussion records as solution knowledge graph

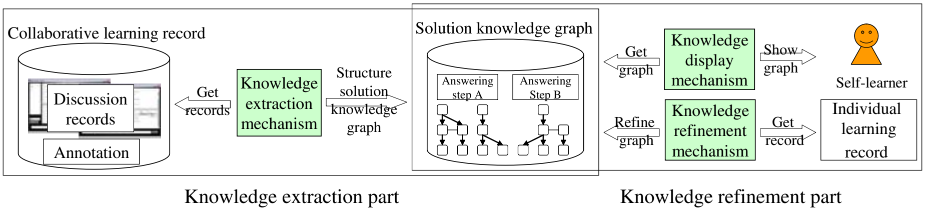

Collaborative learning record
Discussion records
Annotation
Get records
Knowledge extraction mechanism
Structure solution knowledge graph
Solution knowledge graph
Answering step A
Answering Step B
Get graph
Knowledge display mechanism
Show graph
Self-learner
Refine graph
Knowledge refinement mechanism
Get record
Individual learning record
Knowledge extraction part
Knowledge refinement part

Fig. 1. Outline of our system

automatically, and for refining it through its use in individual learnings. Figure 1 illustrates the outline of our system. In the knowledge extraction mechanism, utterance collections are extracted based on the annotations attached by the participants and are classified into answering steps according to their contents. Since annotations attached by the same participants may indicate utterances of the same answering method, utterance collections are classified into answering methods based on the participants who attached the annotations and are structured as a solution knowledge graph. In knowledge display mechanism, solution knowledge graph is shown to self-leaners. In knowledge refinement mechanism, the structure of the solution knowledge graph is changed based on the learning records of self-learners who use the solution knowledge graph as hints for solving the same exercise. By reflecting intention for not only participants but also self-learners, the solution knowledge graph is able to represent utterance collections that are effective for many learners.

3 Solution Knowledge Graph

3.1 Definition of Solution Knowledge Graph

In the solution knowledge graph, useful utterance collections are arranged for answering methods in each answering step. Figure 2 shows the conceptual imagination of the solution knowledge graph. The solution knowledge graph is composed of nodes and links. A node shows a useful utterance collection and contains information on participants/self-learners who used this utterance collection to derive the answer. Nodes at higher positions in the solution knowledge graph correspond to more effective utterance collections, and nodes at lower positions are used by only a few participants/self-learners. Links connect utterance collections with the same answering method. Undirected links are attached to utterance collections whose effectiveness are the same, and directed links indicate that target nodes are supplementary to source nodes.

3.2 Construction of Solution Knowledge Graph

Nodes in the solution knowledge graph express utterance collections. Nodes consist of successive utterances of the same topic that are used for deriving answers. Utterance collections of the same topic is extracted by target utterances of the

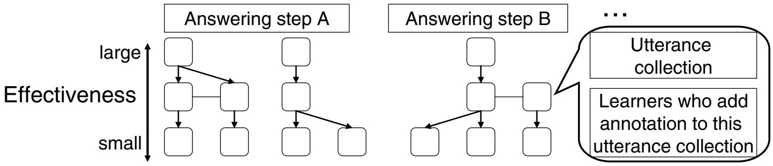


Fig. 2. Solution knowledge graph

utterances that are indicated by participants. Utterances may be derived by their target utterances, and utterances and their target utterances are regarded to express the same topic. So, utterances that are originally derived by the same utterances are gathered and compose utterance collections.

Of all utterance collections, useful ones are used by participants to derive the answer. Since annotations are attached to utterances used for deriving the answer, utterance collections including the utterances to which annotations were attached are extracted as useful utterance collections. The extracted utterance collections form nodes of solution knowledge graph. The number of participants who attached annotations is defined as the degree of effectiveness of the nodes.

The useful utterance collections include knowledge which is necessary for deriving the answer. Therefore, they belong to one of the answering steps. By comparing keywords in each answering step with words included in the utterance collections, nodes are classified into the corresponding answering steps. A participant solves the exercise along one answering method for each answering step. So, if annotations are attached to two utterance collections in one answering step by the same participants, they may belong to the same answering method. *Corresponding rate* is defined as the possible rate that two utterance collections belong to the same answering method.

Corresponding rate between nodes i and j is calculated by Equation 1. N_i is a set of the participants who attached annotations to the utterance collection in node i. When a corresponding rate is more than a threshold, a link is added between the two nodes. When the degrees of effectiveness of nodes i and j are equal, an undirected link is generated between them. If their degrees of effectiveness differ, a directed link is generated from the node with the large effectiveness to the small node.

$$\mathit{corresponding\ rate} = \frac{||N_i \cap N_j||}{||N_i||} \qquad ||N_i|| \geq ||N_j|| \tag{1}$$

3.3 Refinement of Solution Knowledge Graph

The structure of the solution knowledge graph, such as positions of nodes and links, is reconstructed based on learning records of the self-learners. Nodes in the solution knowledge graph that help self-learners to derive answers are counted as useful utterance collections. Therefore, after a self-learner finishes a individual learning, the degree of effectiveness is modified and positions of nodes in the solution knowledge graph are rearranged. Corresponding rates between nodes are also re-calculated based on effective degree that is the number of self-learners who refer to the node to the number of annotations. Equation 2 is the calculation method of corresponding rate in refining the solution knowledge graph. R_i is a set of the self-learners who used the node i. According to the corresponding rate, links of the solution knowledge graph are generated, changed, or deleted.

$$corresponding\,rate = \frac{||(N_i \cup R_i) \cap (N_j \cup R_j)||}{||(N_i \cup R_i)||} \quad ||(N_i \cup R_i)|| \geq ||(N_j \cup R_j)|| \quad (2)$$

4 Prototype System

e construct a prototype system for an individual learning which displays the solution knowledge graph as a knowledge source. The system constructs a solution knowledge graph from discussion records and shows it to a self-learner. Figure 4 shows the interface for the individual learning. Self-learners need to input an exercise number from the combo box and start learning by pushing the start button. When the start button is pushed, an exercise sentence is emerged in the exercise display area, and the solution knowledge graph is drawn in the solution knowledge graph display area. The solution knowledge graph display area is composed of more than one answering step tabs which correspond to each answering step. When an answering step tab is selected, the solution knowledge graph in the corresponding answering step is shown in the solution knowledge graph display area. The rectangles in the solution knowledge graph display area express nodes, and the words in a rectangle represent words that appear more than twice in the corresponding utterance collection.

When the node is clicked, utterances in the selected node are displayed in the discussion display area. When self-learners refer to the utterances that are currently displayed in the discussion display area, the utilization button is pushed.

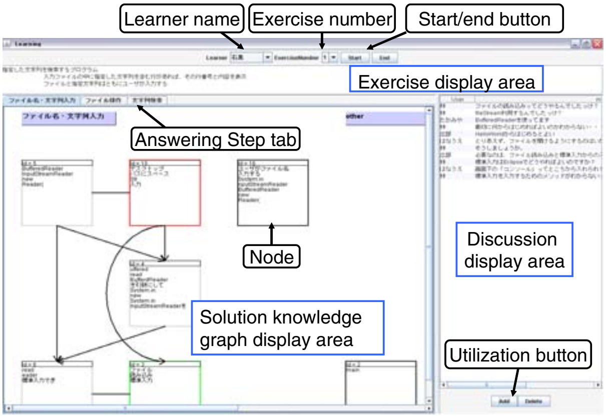


Fig. 3. Interface for individual learning

Then, the system recognizes that the node is used during the individual learning. By pushing the end button, the individual learning is finished and the structure of solution knowledge graph is changed.

5 Experiment

We evaluate the refinement method of a solution knowledge graph. The solution knowledge graph was constructed with discussion records in collaborative learnings of two groups (five and four students in our laboratory). The exercise which participants tackled in the collaborative learnings was "construct a program which retrieves an input string from a file". It is composed of three answering steps, such as "to obtain input string", "to operate a file", and "to retrieve string from the file". The solution knowledge graph was composed of 22 nodes and 10 links. The threshold for generating links was set to 0.5.

Four other students in our laboratory were asked to tackle the same program with the prototype system one by one. They were asked to push the utilization button when referring the node. After each student finished learning, the solution knowledge graph was updated.

Table 1 shows the total number of links changed by four learners. The appropriateness of changed links are evaluated by checking their contents. For example, if links were generated between nodes that indicate the same answering methods, they were regarded as correct links. However, if nodes did not belong to the same answering method, links were attached incorrectly. Eight links were correctly changed, and five links were operated incorrectly. Currently, the influence of one push of the utilization button was large, since the number of students in this experiment was small. Therefore, it is necessary to change the threshold according to the number of self-learners.

On the other hand, Table 2 shows the number of pushing utilization button for nodes in each position in the solution knowledge graph. Layer 1 holds to the most effective nodes and effectiveness of nodes gets smaller as the layer becomes lower. The nodes that were useful for deriving answer were successfully arranged at a higher layer in the solution knowledge graph.

Table 1. Total number of changing links by four self-learners

	correct	incorrect
No. of generated links	2	4
No. of deleted links	6	1
Total No. of changed links	8	5

Table 2. The number of pushing utilization button for nodes in each layer in solution knowledge graph

Position in solution knowledge graph	1st	2nd	3rd	4th	5th	6th
No. of pushing utilization button	10	5	3	4	1	0

6 Conclusion

In this paper, the method for structuring a solution knowledge graph was proposed. The solution knowledge graph was organized by useful utterance collections for deriving answer that were extracted from discussion records. Moreover, it was refined based on records of self-learners who used it. From the experimental result, we confirmed that more useful utterance collections were arranged in higher layer. However, it turned out that links were not generated correctly, especially when the number of participants/self-learners were small. We should reconsider the method for generating links between the nodes of the same answering method by grasping contents of the nodes.

Currently, our system regards utterances collections that many learners referred were more effective. However, all learners do not necessarily need the same knowledge. Effective knowledge maybe different if learners' understanding levels are different. Therefore, a mechanism of recommending utterance collections according to learners' understanding levels should be added.

In addition, our system only allows self-learners to browse the discussion records. In order to organize more effective solution knowledge, the mechanism which modifies discussion records, such as nodes in the solution knowledge graph, need to be developed. For example, a function to comment or correct the utterance collection in nodes is considered.

References

1. Kam, M., Wang, J., Iles, A., Tse, E., Chiu, J., Glaser, D., Tarshish, O., Canny, J.: Livenotes: A System for Cooperative and Augmented Note-Taking in Lectures. In: Proc. of CHI 2005, pp. 531–540 (2005)
2. Dimitracopoulou, A.: Designing Collaborative Learning Systems: Current Trends and Future Research Agend. In: Proc. of CSC 2005, pp. 115–124 (2005)
3. Kayama, M., Okamoto, T.: Knowledge Management Framework for Collaborative Learning Support. In: van Elst, L., Dignum, V., Abecker, A. (eds.) AMKM 2003 Part II. LNCS, vol. 2926, pp. 107–117. Springer, Heidelberg (2004)
4. Goodman, B., Geier, M., Haverty, L., Linton, F., Mcready, R.: A Framework for Asynchronous Collaborative Learning and Problem Solving. In: Proc. of AIED 2001, pp. 188–199 (2001)
5. Adams, P.H., Martell, C.H.: Topic Detection and Extraction in Chat. In: Proc. of ICSC 2008, pp. 581–588 (2008)
6. Okazaki, N., Matsuo, Y., Matsumura, N., Ishizuka, M.: Sentence Extraction by Spreading Activation through Sentence Similarity. IEICE Trans. on Information and Systems E86D(9), 1686–1694 (2003)
7. Kakehi, M., Kojiri, T., Watanabe, T., Yamada, T., Iwata, T.: Organization of Discussion Knowledge Graph from Collaborative Learning Record. In: Diekert, V., Volkov, M.V., Voronkov, A. (eds.) CSR 2007. LNCS (LNAI), vol. 4649, pp. 600–607. Springer, Heidelberg (2007)

Implementation of Wireless Sensor System and Interface for Agricultural Use

Kenji Obata[1], Takahiro Masui[1], Hiroshi Mineno[2], and Tadanori Mizuno[3]

[1] Graduate School of Informatics, Shizuoka University
[2] Department of Computer Science, Shizuoka University
[3] Graduate School of Science and Technology, Shizuoka University
3-5-1 Jyohoku, Naka-ku, Hamamatsu, Shizuoka, 432-8011, Japan
{ova,masui}@mizulab.net,
{mineno,mizuno}@inf.shizuoka.ac.jp

Abstract. Problems involving agricultural know-how can be addressed with the use of IT. For example, IT can reduce the risk that know-how may be lost due to the increasing age of agricultural workers. Also valuable fruits which require sensitive environmental control can be monitored with IT. Data collection, collation and storage will enable us to convert tacit knowledge into formalized algorithms. We made a remote information sensing system using a sensor board made in cooperation with Renesas Solutions Corporation. Furthermore, we had installed our sensing system in a melon hothouse, in cooperation with the Prefectural Research Institute. We set up an interface enabling access to data and photographs using a browser. Using this interface, we expect that farmers will be able to transfer tacit knowledge into formalized information.

Keywords: Wireless sensor network, Remote sensing, Agricultural sensor system, User interface.

1 Introduction

The agricultural demographic is changing; the median age of farmers increases while their numbers diminish. Precious agricultural know-how is in danger of being lost [1]. Moreover, cash crops, like hot-house melons (used in the present project) command prices upwards of $200 per fruit. It is important to maintain constant conditions in order to produce a uniformly high quality product. IT has potential for solving both of these problems. The painstaking procedure of drawing out the tacit knowledge of expert farmers can be facilitated with a data gathering protocol utilizing sensors, sensor board, and computer interface, enabling the system to collate and store a variety of searchable data [2]. This will enable us, on the one hand, to gather the expert knowledge of aging farmers for preservation and dissemination. The information can be utilized by anyone. Also we can facilitate the consistent control of the delicate environment with sensors monitoring such factors as temperature, atmospheric moisture, ground moisture, and light and so ensure a reliably high grade of produce.

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 572–579, 2009.

We developed a remote information sensing system using a sensor board made in cooperation with Renesas Solutions Corporation [3]. Sensor networks are critical tools for measuring and understanding the complex interactive dynamics of natural systems [4], and help promote new ways of agricultural management [5]. The cost of sensor technology is expected to fall in the near future while utility of IT will continue to develop for various agricultural applications [6][7]. We are developing the protocol for initiating the process of accumulating the massive store of information required to make agricultural knowledge available to everyone.

2 Remote Sensing System

2.1 Our Sensor Boards

This research concerns the construction of a sensor network utilizing a ZigBee node produced by Renesas Solutions Corporation. The Renesas node is used with a second board to make our sensor system. The first is the battery operated Renesas sensor board in Fig. 1 with three kinds of sensors: motion, temperature, and light. It can be operated either with 4 AAA batteries or an AC adaptor. This board is able to collect sensor values. This board, used as the sensor node, is the topic of this paper.

The other board is a ZigBee evaluation board shown in Fig. 2. It has no sensors and batteries, and therefore a power supply cable is required. This board is equipped with RS-232C interface which can be connected with a PC. The board can be used in three ways, one board each: the first as a ZigBee router, the second as a coordinator and the third as a sink node.

2.2 Basic Topology

The basic topology of the system consisting of these boards is shown in Fig. 3. The network topology shows the structure of the procedure. Each sensor is connected to the ZigBee router. The router channels the data from the sensors to the sink node. The sensors must be placed within the transmission range of the router. If the router is placed within the range of a sequential router, it is possible to transmit multi-hop data. Thus a continuous flow of data is sent to the sink node.

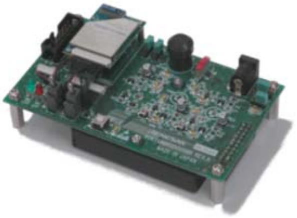


Fig. 1. Sensor board

Fig. 2. ZigBee evaluation board

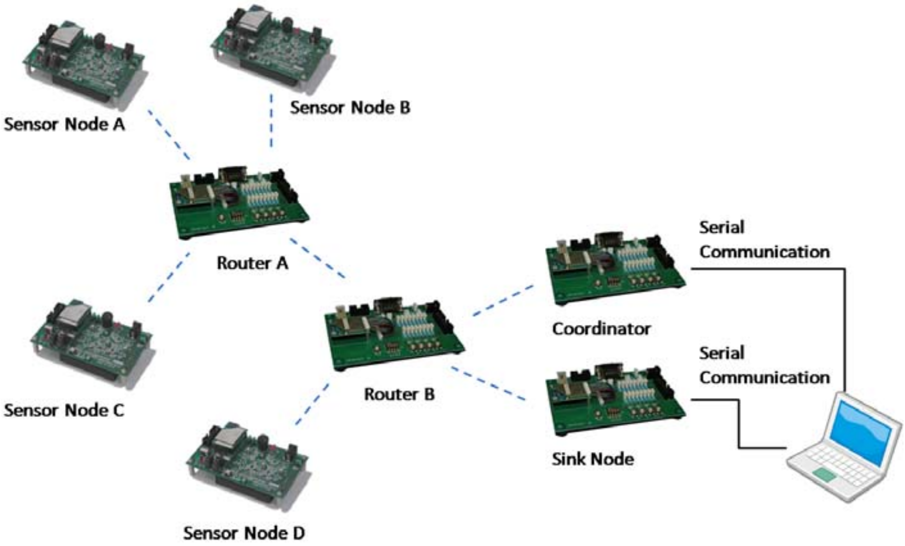


Fig. 3. Basic topology of sensing system

2.3 Sensing Data Gathering

The data received by the RS-232C serial port is then transferred to the PC. The information is then recorded in the data base (Fig. 4). The data received by serial communication is recorded in the Gateway HDD in a temporary file. Data is accumulated and transferred as a transaction at regular intervals to the PostgreSQL server in our laboratory. Because of the volume of the continual stream of data received from the sensors, once a day the program compiles the incoming data packets in searchable tables formatted with time, ID, and sensor readings categories.

2.4 Camera Data Gathering

Moreover, in response to the client user's request for more detailed visual information concerning data received from the sensors, photographic information is also recorded. A separate ZigBee network gathers photographs from network

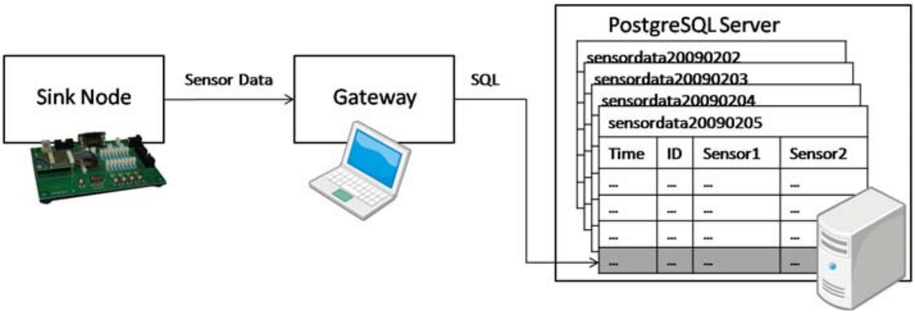


Fig. 4. Block diagram of sensor data gathering

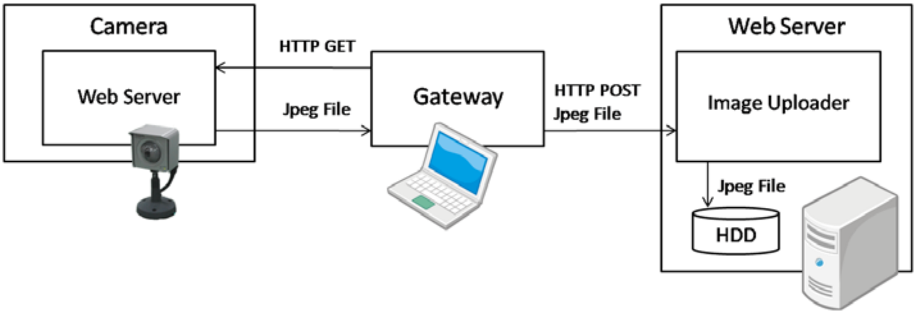


Fig. 5. Brock diagram of camera data

cameras at 15 minute intervals and transfers them by LAN using HTTP protocol to a designated server (Fig. 5).

The routing is arranged as follows: The network camera is equipped with a Web server. The Gateway accesses the network camera using HTTP protocol and receives the Jpeg files which it then relays by HTTP post to the server where a PHP program for image uploading transfers the Jpeg to the designated HDD for graphic images. The separately maintained data from the sensors and the photographic images can be collated later for viewing. It is beneficial to monitor the atmospheric conditions of the hothouse which affect the produce (i.e. melons) by having a sensor system to monitor variables such as temperature, moisture, and light and collate them for later analysis.

3 Report of Melon House Experiment

We prepared the above described system of sensors, cameras, and servers for an experiment in remote sensing we placed the system in the melon hothouse, and connected it with the sensor node inside the building (Fig. 6). We placed the Sink Node and Gateway in a separate unmanned observation room 15 meters from the hothouse.

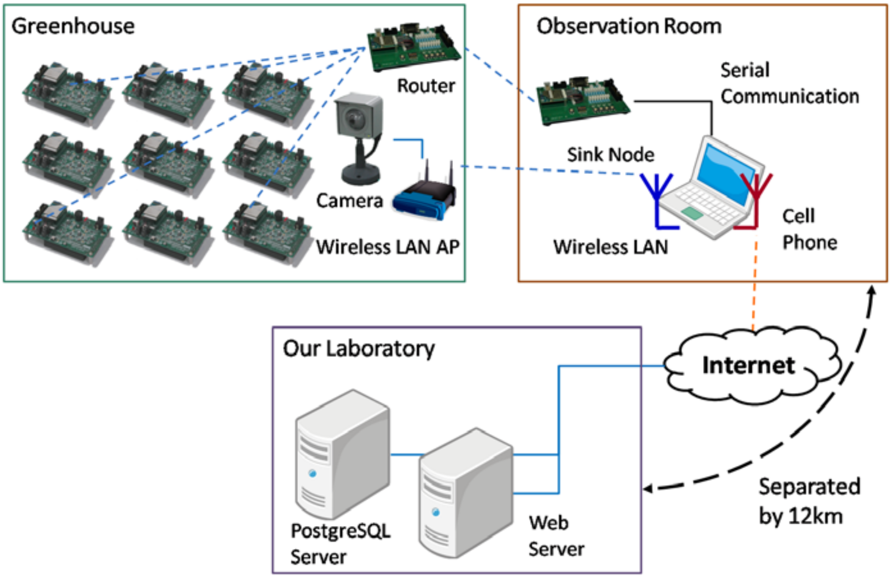


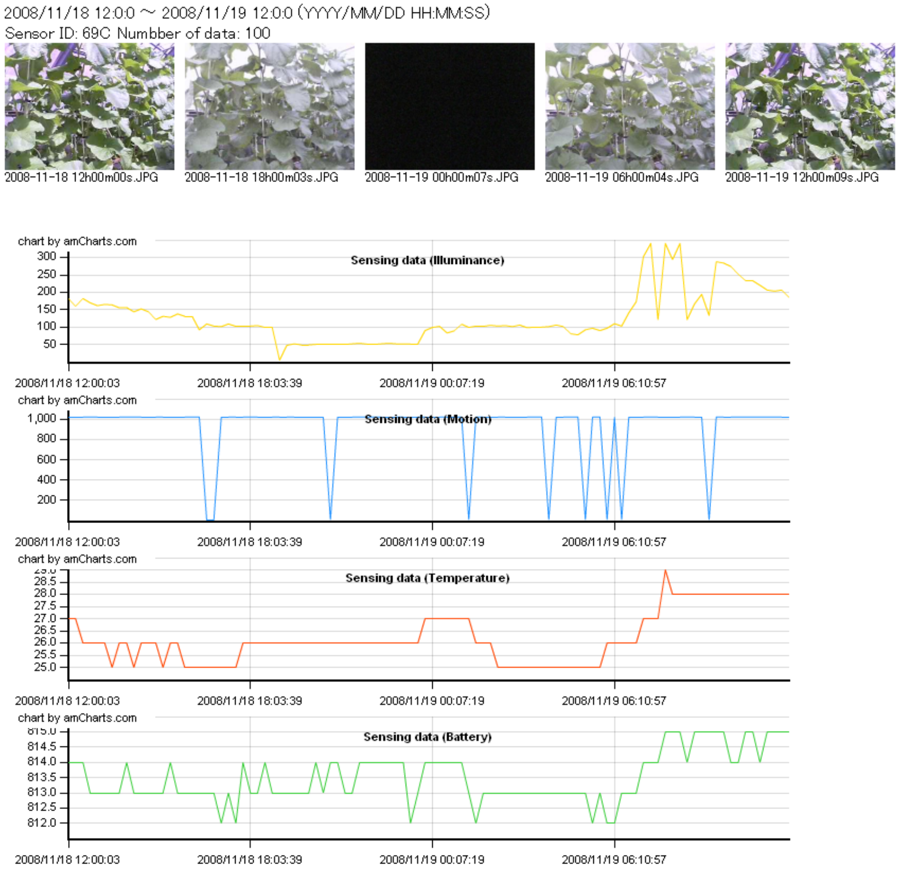
Fig. 6. Sensing system set in hothouse and observation room

In the melon house we placed the router in position so as to gather the information from all the nodes. Information is communicated by ZigBee between sensor node and sink node as described above. Because the camera is equipped with LAN cable we connected it to a wireless LAN access point inside the hothouse which transmitted data to the observation room. We made provisional use of cell phone to provide the internet access for the Gateway PC. We used this system to transfer the data from sensors and the photographic images from the network cameras to the laboratory.

3.1 Interface

We made an interface to put the information thus gathered by sensors and cameras into a readable format. It is possible to access this interface from a Web browser (Fig. 7). Sensor information during a day is displayed in this figure. The photograph is displayed in upper part of the page, and values of the light, motion, temperature, and voltage of equipped batteries have been graphed. At the time of accessing the interface to view selected records, the user inputs the ID for the selected sensor, a start and finish term and the number of records required, distributed evenly in the time period.

Because of the huge volume of information from the sensors, when all the information is downloaded, the infrastructure between the Web server and the PostgreSQL server is overloaded. As a solution for this problem, we implemented the function inside the PostgreSQL server to request the required amount of data (Fig. 8). The ID, term, and number of records of the desired data is transmitted



2008/11/18 12:0:0 ～ 2008/11/19 12:0:0 (YYYY/MM/DD HH:MM:SS)
Sensor ID: 69C Numbber of data: 100

Fig. 7. Physical appearance of Web Interface

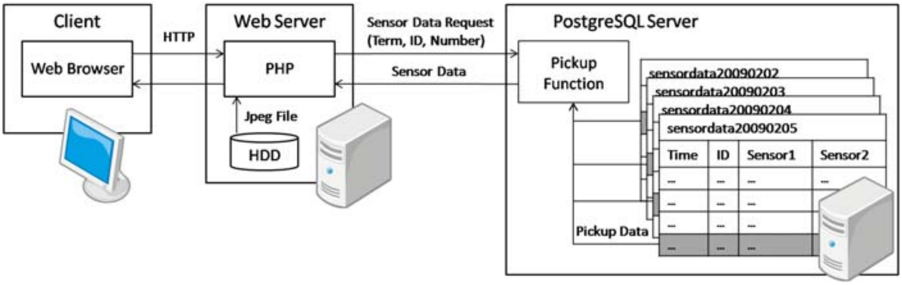


Fig. 8. Web interface block diagram

to the PostgreSQL server from the Web server PHP. The selection function, using a simple algorithm, gathers the requested information from the tables on the PostgreSQL database according to timestamp and sends them back to the PHP. The PHP makes an XML file from the PostgreSQL data and reads it with a graphic program. Then the PHP gets the saved Jpeg photographs on the HDD for the required period and sends them together to the client's browser.

4 Results

We conducted this experiment for 10 days 24 hours per day, and were able to successfully retrieve requested information. We tested the system by requesting reports from the laboratory. We monitored the equipment 8 times a day for proper functioning from the laboratory. System failure occurred only once. The cell phone connection failed and needed to be reconnected. We tested data sampling protocol (sensor ID, term, number of records) everyday. It was largely successful but the problems related to the cell phone must be resolved, whether by compressing data or replacing the cell phone with some other broadcast equipment. This research is still in progress and our goal here is to report the purpose and structure of the system.

5 Summary

In this paper, we have reported the experiment we conducted for a melon hot house with a remote sensing system that we constructed using a Renesas board, plus sensors, cameras, and servers. We expect that hereafter agricultural engineers will refine this system utilizing sensors for light, temperature, ground and air moisture, as well as potentially unlimited other factors. Access to sensor information will promote more reliable results. The reliable results will inspire the trust of the agricultural workers in the relationship between the information provided by sensors and the produce. When they understand the relevance of IT to their livelihood, agricultural producers will work to formalize their tacit knowledge. It is thought that by using formalized knowledge, agricultural best-practice can be widely disseminated. This will be a contribution to agricultural development.

Acknowledgments. This research was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Young Scientists (A), 21680007, 2009.

References

1. Abdon, B.R., Raab, R.T.: Knowledge Sharing and Distance Learning for Sustainable Agriculture in the Asia-Pacific Region: the Role of the Internet. Plant Production Science 8, 298–307 (2005)

2. Baalen, P.V., Ruwaard, J.B., Heck, E.V.: Knowledge Sharing in an Emerging Network of Practice: The Role of a Knowledge Portal. European Management Journal 23, 300–314 (2005)
3. Renesas Solutions Corporation, `http://www.rso.renesas.com/`
4. Wark, T., Corke, P., Pavan, S., Klingbeil, L., Guo, Y., Crossman, C., Valencia, P., Swain, D., Hurley, G.B.: Transforming Agriculture through Pervasive Wireless Sensor Networks. Pervasive Computing, IEEE 6, 50–57 (2007)
5. Kitchen, N.R.: Emerging technologies for real-time and integrated agriculture decisions. Computers and Electronics in Agriculture 61, 1–3 (2008)
6. Wang, N., Zhang, N., Wang, M.: Wireless sensors in agriculture and food industry–Recent development and future perspective. Computers and Electronics in Agriculture 50, 1–14 (2006)
7. Hebel, M.A.: Meeting Wide-Area Agricultural Data Acquisition and Control Challenges through ZigBee Wireless Network Technology. In: Computers in Agriculture and Natural Resources, 4th World Congress Conference, pp. 234–239 (2006)

Algorithms for Extracting Topic across Different Types of Documents

Shoichi Nakamura[1], Saori Chiba[1], Hirokazu Shirai[2], Hiroaki Kaminaga[1],
Setsuo Yokoyama[2], and Youzou Miyadera[2]

[1] Fukushima University, Department of Computer Science and Mathematics,
Kanayagawa 1, Fukushima, 960-1296 Japan
nakamura@sss.fukushima-u.ac.jp
[2] Tokyo Gakugei University, Division of Natural Science,
4-1-1, Nukui-Kita, Koganei, Tokyo, 148-8501 Japan
{miyadera,yokoyama}@u-gakugei.ac.jp

Abstract. Clever management of the various types of documents used in intelligent activities and their efficient utilization are important. However, most available methods target only a single type of document (e-mails, Web pages, etc.). A more promising approach is topic-centered document management. Algorithms are described for extracting topics across various of types of documents. Moreover, a topic-centered document management system is described that is based on grouping by topics.

1 Introduction

Efficient discovery and utilization of useful information from various types of documents is important in intelligent activities, e.g., research activities and cooperative software development in a network environment. However, the number of documents related to such activities increases exponentially as the activities progress. This makes it harder and harder to identify the useful documents. There is thus a strong need for ways to support the management and utilization of documents in accordance with the how the information is to be used.

Although there has been research on topic extraction [1][2][3] and document clustering [4][5][6][7][8], each of these research projects targeted only one type of document. As a result, the existing methods are unsuitable for topic extraction and clustering across different types of documents. Thus, topic extraction independent of document type and topic-centered management of documents are needed to achieve clever document management in accordance with how the information is to be used.

This paper describes algorithms that have been developed for topic extraction across document types and methods that support document management based on the topic.

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 580–590, 2009.

2 Problems in Document Management Related to Purpose of Use and Support Policies

2.1 Research Targets

The target of this research was intelligent activities in a network environment such as research activities, exploratory learning, and cooperative software development. In these activities, it is important to discover and efficiently utilize the desired documents in accordance with how the information is to be used. Since users accumulate various types of documents and in increasing quantities as activities progress, it is often difficult to identify the useful documents from the many documents accumulated.

The search for a desired document is enforced by focusing on the purposes of use. The topics hidden in the numerous documents are used here as indicators of the purpose of use. However, the elements of the documents differ with document type (e-mail, Web page, and so on). Consequently, topic extraction is harder when various types of documents are involved.

Moreover, the software applications differ with the document type; for example, mail client software is used for e-mail, and a Web browser is used for Web pages. Therefore, users can become bewildered as they come and go among various applications. Accordingly, it can be difficult for users to manage different types of documents while relating them sufficiently. This can make it difficult to grasp the progress of a project and the relationships among documents.

Consequently, support is needed for clever document management that enables users to locate and utilize useful documents regardless of the document type and application. This research targets document management by people working together such as on a laboratory research project. We assume that all of them use e-mail and use several types of documents.

2.2 Related Research

Many methods for topic extraction and document clustering have been reported. Hamasaki et al. described a method for discovering networks of common topics from bookmarks [1]. This method uses the hierarchical structure of bookmarks and identifies potentially useful pages by investigating the communities of topics among users. Sekiguchi et al. developed a method for extracting topics that uses the characteristics of utterances in weblogs [2]. Moreover, a trial in which networks of human relationships were extracted from information on the Web [9] has been reported.

The research on document clustering includes the work by Yanai et al. on automatic image clustering [4]. They developed a method for clustering natural images from the real world on the basis of learning from images gathered from the Web. Iyama et al. developed a system for clustering Web pages on the basis of their characteristics and for providing the results to users [5]. These clustering methods are unable to extract the topic across various types of documents and to cluster different types of documents since they each target only one type of document.

Systems that support document management and sharing have also been developed. Sano et al. described an information sharing system [10] that analyses the bookmarks of many users and uses the results to recommend a URI for users with similar interests. An interesting direction for research is the support of not only topic extraction but also of information sharing based on it. Moreover, Gmail [11], Google's e-mail program, is a strong tool that groups e-mails automatically on the basis of their reply relationships and supports management of e-mails on basis of the extracted groups. Nevertheless, most available systems support only one type of document. Therefore, managing different types of document efficiently is difficult. Furthermore, support for identifying the relationships among documents has not been investigated enough although such an identification would also contribute to understanding the work processes.

2.3 Issues and Support Policies

From the above discussion, we can identify three issues that need to be addressed in order to realize support for document management and utilization in accordance with how the information contained in them will be used.

1. Difficulty of topic extraction from different types of documents.
2. Difficulty of managing different types of documents while relating them.
3. Difficulty of understanding the relationships among documents.

To resolve these issues, algorithms have been developed for extracting the topic across a variety of document types. Moreover, an adaptive document management system has been developed that constructs its own functions and interfaces and changes them in accordance with the target types of documents and the purposes for which the information they contain will be used. This system has a function to assist understanding of work processes, document transitions, and their update circumstances by visualizing the relationships among documents related to a topic.

Here, topic extraction methods can be classified roughly into two types: 1) those that use a natural language based approach to either extract the contents of documents in a semantically constrained way or summarize them (e.g., [12]); 2) those that extract groups of documents that correspond to a topic (e.g., [13]). In this research, the latter approach was used as it aims to realize clever management and utilization of documents in accordance with how the information they contain will be used. The aim is to realize clever management by developing a topic-centered adaptive document management system (as described in Section 5).

3 Algorithm

3.1 Overview

A topic is represented by a group of documents corresponding to that topic. Each topic is considered to have various attributes (members, keywords, activity period, and so on). The set of these attributes, which expresses the feature of the topic, is called the "topic object." When a new document is acquired, the candidate topic to which it belongs is estimated on the basis of the coincidence rates, which are calculated by comparing the attributes of the new document with those of the existing topic objects. The coincidence rate is basically the general state of whether two attributes agree or not. The calculation methods differ for each combination of attributes, and they have been defined temporarily on the basis of an initial investigation.

Figure 1 illustrates the topic extraction algorithm. First, e-mails are grouped on the basis of the reply relationships. The resulting e-mail groups are used as a basic set of documents that expresses the topic. If there is more than one topic (group of documents), the attributes shown in Table 1 are extracted from each group. Each topic thereby acquires a set of elements that express its features, i.e., the topic object.

When a new document is acquired, the attributes specified beforehand in accordance with the types of documents (Table 1) are initially extracted. For the present time, e-mails, shared bookmarks, and PDF files are supported. If there is more than one topic, the relationship degree, which expresses the strength of the relationship between the new document and each topic, is calculated by comparing the attributes of both. If the calculated relationship degree exceeds a specified threshold, the topic with the maximum relationship degree is selected as the candidate topic for the new document. Conversely, if the relationship degrees for all topics fall below the threshold, the new document is judged to not belong to any topic.

Furthermore, if more than a specified number of e-mails belong to no topic before the calculation of the relationship degree, grouping of the e-mails is repeated. In this manner, new topics appearing after the initial grouping are handled.

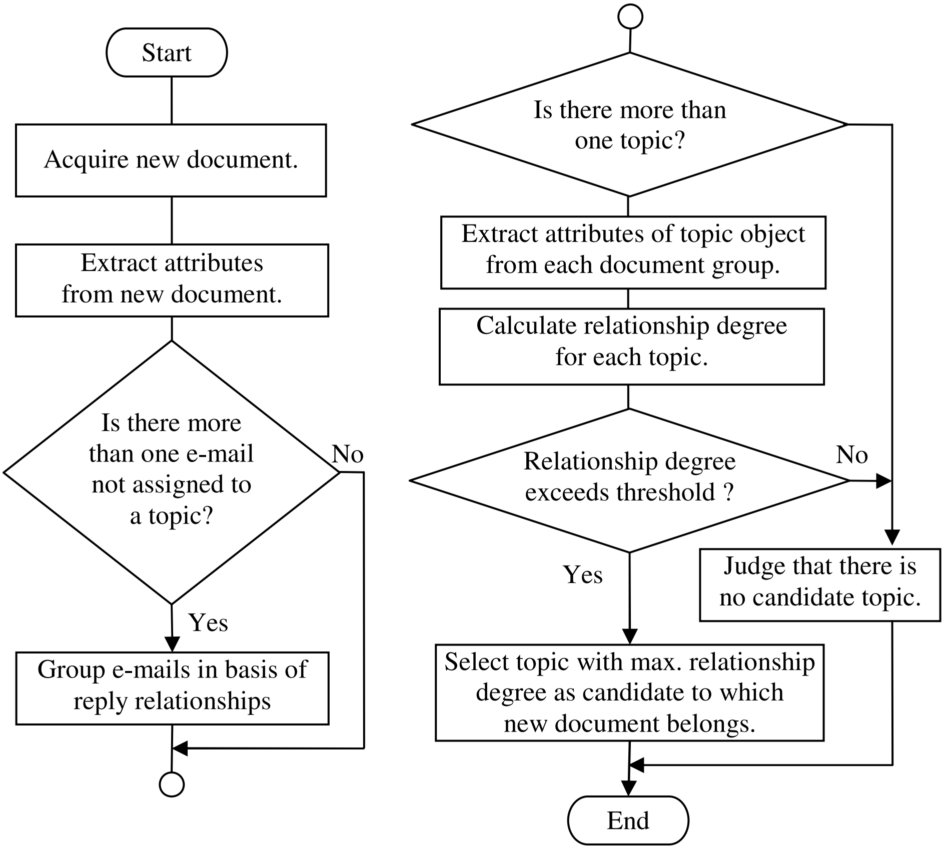
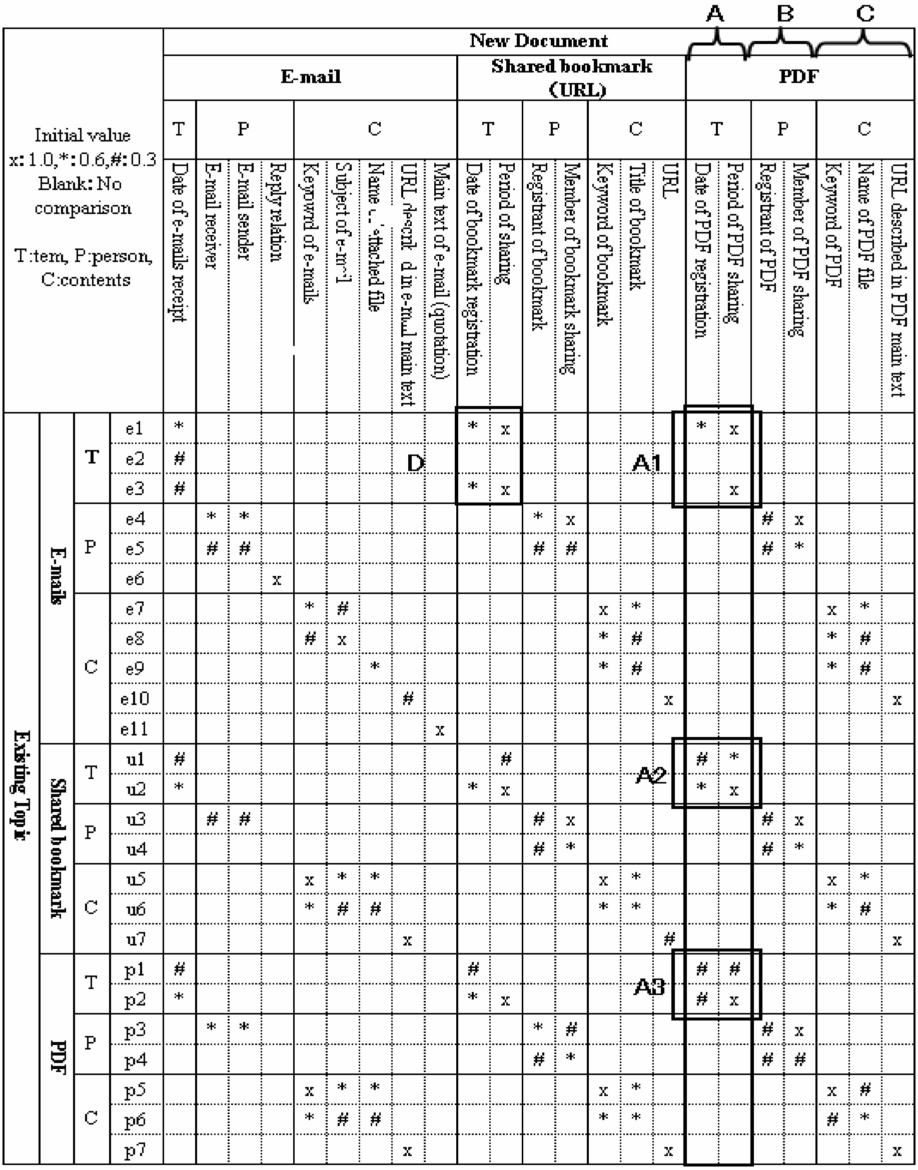

Start
Acquire new document.
Extract attributes from new document.
Is there more than one e-mail not assigned to a topic?
No
Yes
Group e-mails in basis of reply relationships
Is there more than one topic?
Extract attributes of topic object from each document group.
Calculate relationship degree for each topic.
Relationship degree exceeds threshold ?
No
Yes
Judge that there is no candidate topic.
Select topic with max. relationship degree as candidate to which new document belongs.
End

Fig. 1. Topic extraction algorithm

Table 1. Relationships among attributes of topic objects



																				A		B		C		
Initial value x:1.0,*:0.6,#:0.3 Blank: No comparison T:tem, P:person, C:contents				New Document																						
				E-mail									Shared bookmark (URL)							PDF						
				T	P			C					T		P		C			T		P		C		
				Date of e-mails receipt	E-mail receiver	E-mail sender	Reply relation	Keyword of e-mails	Subject of e-mail	Name of attached file	URL described in e-mail main text	Main text of e-mail (quotation)	Date of bookmark registration	Period of sharing	Registrant of bookmark	Member of bookmark sharing	Keyword of bookmark	Title of bookmark	URL	Date of PDF registration	Period of PDF sharing	Registrant of PDF	Member of PDF sharing	Keyword of PDF	Name of PDF file	URL described in PDF main text
Existing Topic	E-mails	T	e1	*									*	x						*	x					
			e2	#							D							A1								
			e3	#									*	x							x					
		P	e4		*	*									*	x						#	x			
			e5		#	#									#	#						#	*			
			e6				x																			
		C	e7					*	#								x	*						x	*	
			e8					#	x								*	#						*	#	
			e9							*							*	#						*	#	
			e10								#								x							x
			e11									x														
	Shared bookmark	T	u1	#										#				A2		#	*					
			u2	*									*	x						*	x					
		P	u3		#	#									#	x						#	x			
			u4												#	*						#	*			
		C	u5					x	*	*							x	*						x	*	
			u6					*	#	#							*	*						*	#	
			u7								x								#							x
	PDF	T	p1	#									#					A3		#	#					
			p2	*									*	x						#	x					
		P	p3		*	*									*	#						#	x			
			p4												#	*						#	#			
		C	p5					x	*	*							x	*						x	#	
			p6					*	#	#							*	*						#	*	
			p7								x								x							x

e1:Period of e-mails, **e2**:Average of e-mail reply, **e3**:Active period of e-mail, **e4**:Member of e-mail, **e5**:Key person, **e6**:Reply relationships, **e7**:Keyword of e-mail, **e8**:Subject of e-mail, **e9**:Name of attached file, **e10**:URLs in e-mail main text, **e11**:Main text of e-mail (quotation), **u1**:Bookmark registration date, **u2**: Bookmark sharing period, **u3**:Member of bookmark sharing, **u4**:Registrant of bookmark, **u5**:Keyword of bookmark, **u6**:Title of bookmark, **u7**:URL, **p1**:Date of PDF registration, **p2**:Period of PDF registration, **p3**:Member of PDF sharing, **p4**:Registrant of PDF, **p5**:Keyword of PDF, **p6**:Name of PDF file, **p7**:URL described in PDF main text.

3.2 E-mail Grouping

The focus here is on e-mail as this is the most common type of document and most everyone uses it. Groups of e-mails are used to express the basis of topics. Each e-mail has a unique ID called a Message-ID. Moreover, either “In-reply-to” or “Reference” information is added to the header of an e-mail sent in reply to a previous e-mail.

“In-reply-to” indicates the original e-mail by its Message-ID. “Reference” expresses a sequence of e-mails included in either previously sent e-mails or replied-to ones by their Message-IDs. Generally, at least one of these information fields is added into to the header of a reply e-mail. Here, e-mails are grouped by analyzing them.

3.3 Extraction of Topic Object

Several attributes are extracted from each group of e-mails created by analyzing the coincidence of the reply relationships. The extracted attributes are used as factors of the topic object as an initial step.

- **Topic period**: For each group, extract earliest and latest e-mail and use their creation dates as dates when the topic started and ended.
- **Average reply interval**: For each group, calculate average time from when e-mail was sent to when a reply to it was sent.
- **Keyword**: For each group, extract the top five characteristic words from the main texts of the e-mails.
- **Member**: For each group, extract the sets of senders and receivers.
- **Attached file**: For each group, extract the set of attached files.
- **Subject**: For each group, extract the set of subjects.
- **Reply Info.**: For each group, extract the “Message-ID” and either the “Reference” or “In-reply-to” from the e-mail headers.
- **URL**: For each group, extract the URLs from the main texts of the e-mails.

3.4 Calculation of Relationship Degree

This section describes the method used to calculate the relationship degree used for judging the topic to which a new document belongs.

When a new document is acquired, the specified attributes are extracted. These typical attributes were selected on the basis of previous investigation and experience. Then, the relationship degree is calculated by comparing the extracted attributes and the factors of each topic object. However, there are no comparable attributes between the new document and topic object when the document type of new document is not included in the target object. For instance, if a bookmark is registered immediately after initial grouping comparison is impossible since a topic object consists only of e-mails. This problem was solved by investigating the relationships among the factors of the various types of documents that went into building the topic object and then specifying the initial relationships, as shown in Table 1.

The main attributes that characterize the topics vary by topic and user. As a countermeasure against this problem, three meta-divisions were introduced for the topic object factors: term, person, and contents (Table 1).

The relationship degree of document *i* to document *j*, *rel(i,j)*, is calculated using equation (1). First, the coincidence rates between the combinations of factors are calculated. The target combinations and their initial weights are specified as shown in Table 1. The relationship degree is expressed as a value that is more than 0 and less than 1. The sum of the coincidence rates between the new document and the existing topics is finally calculated using normalized weights based on preliminary experiments for every area of the combinations of meta-divisions and the types of documents (e.g., when type of new document is PDF, A1 in Table 1).

Then, the calculated sum is multiplied by the ratio of the number of comparison targets to the number of all documents in the topic. The same calculation is done for areas A2 and A3 as well. The sum of these three values is the coincidence rate for each meta-division (e.g., A in Table 1): $agr_{ter}(i,j)$, $agr_{per}(i,j)$, $agr_{con}(i,j)$.

Finally, relationship degree *rel(i, j)* is obtained by summing the coincidence rates of the three meta-divisions multiplied with their respective weights: $w_{ter}(j)$, $w_{per}(j)$ $w_{con}(j)$ Although documents could belong to more than one topic, it is assumed that each document has only one candidate topic.

$$rel(i,j) = w_{ter}(j) \cdot agr_{ter}(i,j) + w_{per}(j) \cdot agr_{per}(i,j) + w_{con}(j) \cdot agr_{con}(i,j)$$

$$1 \geq w_{ter}(j), w_{per}(j), w_{con}(j) \geq 0,$$
$$w_{ter}(j) + w_{per}(j) + w_{con}(j) = 1 \quad (1)$$

4 Methods for Updating Relationships

4.1 Overview

Document classification differs by user and by circumstance. In one case, a user might cluster documents on the basis of their contents such as when documents related to a research theme are gathered. In another case, a user might arrange documents on the basis of the creator or receiver such as when e-mails received from and/or sent to a particular person are gathered. In a third case, a user might manage documents on the basis of time such as when documents created during a certain period are gathered.

An algorithm for topic extraction should be able to handle these various cases. The method described in this section can be used to update the relationships among the attributes of the topic object and meta-divisions of the attributes.

4.2 Method for Updating Relationships among Attributes of Topic Object

As mentioned above, relationships among attributes of topic objects differ with the topic and the user. Therefore, in the method described here, updating is done topic by topic.

Consider this example: an attribute is selected from the "term" factors for e-mail as the existing document and an attribute is selected from the "term" factors of bookmarks as the new document (D in Table 1). Initially, two types of attributes are selected. If the combination of the selected two attributes is marked (weighted) in Table 1, the relationship between those two attributes is evaluated in the same way as the relationship degree is calculated. Next, how much the attributes of the term divisions (Registration date, Sharing period) of all bookmarks belonging to the same topic coincide with those of e-mails (Period of e-mails, Active period) is evaluated. The average of that rate becomes the new relationship value. Specifically, the following coincidence rates are calculated, except for the blank cells in Table 1.

- Period of e-mails includes bookmark registration date or not (Includes: 1, Not included: 0).
- Rate of period of e-mails includes bookmark sharing period (number of overlapping days between period of e-mails and bookmark sharing period / total number of days for period of e-mails).
- A bookmark is registered during active period of e-mails (Registered: 1, Not registered: 0).
- Rate of active period of e-mails includes bookmark sharing period (number of overlapping days between active period of e-mails and bookmark sharing period / total number of days for active period of e-mails).

That is, evaluation in this updating examines the rate of influence that coincidence between each factor of topic object affects the belonging of documents to the topic, while calculation of relationship degree for a new document is used to judge the candidate topic to which the new document belongs.

For combination of other types of documents, the relationships among them are updated in the same way. The timing for the updating is determined on the basis of the date of the last updating, the number of accumulated documents, and so on.

4.3 Method for Updating Relationships among Meta-divisions

The updating of the relationships among meta-divisions of the attributes of topic objects involves term, person, and contents. This updating is done for each topic and user, the same in updating the relationships among attributes. Specifically, updating is done in accordance with the following procedure.

(1) Select one document from all documents belonging to the target topic.

(2) Calculate the coincidence between the attributes of the selected document and those of topic object same to calculation of relationship degree (more than 0 and less than 1).

(3) Calculate the average coincidences for the three meta-divisions (more than 0 and less than 1).

(4) Repeat steps (1) to (3) for all documents and then calculate the averages for the three meta-divisions.

(5) Replace the rate for the averages of the three meta-divisions with the new calculated averages.

This updating and replacement should improve the accuracy of topic extraction across different of types of documents and contribute to achieving topic-centered document management. The timing and frequency of the updating should be investigated since it requires calculation costs.

5 Topic-Centered Document Management System

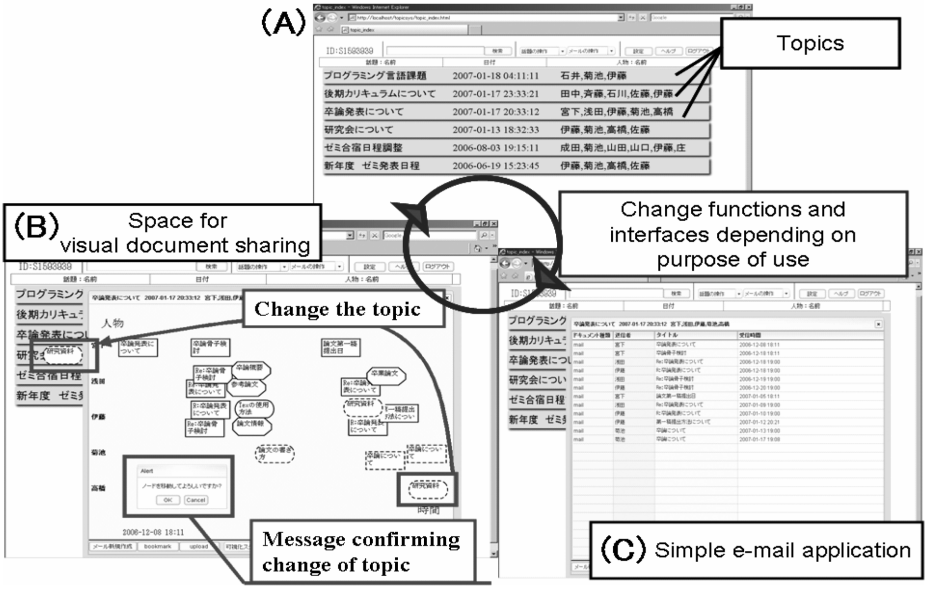
This research aims at developing a topic extraction method that can handle a mixture of different-types electronic documents and at developing an electronic document management system based on grouping by topics. This section describes the outline and basic idea of such a system.

In this system, the elements indicating the topic are automatically extracted from electronic documents like e-mails, and the extracted topics are used as a basic unit for various operations in document management. For example, at the time of system startup, as shown in Figure 2(A), rather than individual documents, topics are presented by listing characteristic words extracted from documents or elements of members.

The management area corresponding to the selected topic is shown, and then several types of documents belonging to the topic are displayed as a node with a different shape and a different color (Figure 2(B)). When a newly arrived document is judged to belong to a topic, it is shown as a special node in the management area of that topic. The user can change the topic judged by the system through interactive operation using a semi-automatic system.

Moreover, to help the user grasp the relationships among documents corresponding to various viewpoints, the visual presentation of relationships among documents is done considering various factors (e.g., time, creator, types of documents). This visualization can also contribute to grasp processes of works and to share them.

Furthermore, there is adaptive construction of functions and interfaces corresponding to the situation. For example, the system sometimes works as simple e-mail


(A)
Topics
(B) Space for visual document sharing
Change functions and interfaces depending on purpose of use
Change the topic
Message confirming change of topic
(C) Simple e-mail application

ID:S1593939

話題：名前	日付	人物：名前
プログラミング言語課題	2007-01-18 04:11:11	石井,菊池,伊藤
後期カリキュラムについて	2007-01-17 23:33:21	田中,斉藤,石川,佐藤,伊藤
卒論発表について	2007-01-17 20:33:12	宮下,浅田,伊藤,菊池,高橋
研究会について	2007-01-13 18:32:33	伊藤,菊池,高橋,佐藤
ゼミ合宿日程調整	2006-08-03 19:15:11	成田,菊池,山田,山口,伊藤,庄
新年度 ゼミ発表日程	2006-06-19 15:23:45	伊藤,菊池,高橋,佐藤

Fig. 2. Topic-centered document management system

software (Figure 2(C)) and it other times evolves into an integrated shared environment in which documents of various types are treated (Figure 2(B)).

In conventional styles, users are often bewildered as they come and go among various applications because available applications depended on types of documents. The system described here enables new relationship between systems and a user so that systems adapt themselves to the needs of the user.

6 Conclusions

This paper describes algorithms for topic extraction across document types using topic object. In addition to details of the algorithm for judging topic to which new document belongs, methods for updating relationships among attributes of topic object and among meta-divisions of the attributes are also described. Finally, an adaptive system to realize topic-centered document management is proposed.

Future work includes designing and developing a prototype system and evaluating the topic extraction algorithm.

Acknowledgments

This work was supported in part by the Ministry of Education, Culture, Sports, Science and Technology, Japan and by Japan Society for the Promotion of Science under Grant-in-Aid for Young Scientists (B) (No.20700632), Grant-in-Aid for Scientific Research (B) (No.21300304) (No. 19300272) respectively.

References

1. Hamasaki, M., Takeda, H., Matsuzuka, T., Taniguchi, Y., Kono, Y., Kidode, M.: A Method of Discovery of Shared Topic Networks among People from WWW Bookmarks and Its Evaluations. Trans. JSAI 17(3), 276–284 (2002) (in Japanese)
2. Sekiguchi, Y., Kawashima, H., Okuda, H., Oku, M.: Topic Detection from Blog Documents Using Bloggers' Interest. DBSJ Letters 5(1), 9–12 (2006)
3. Lamping, J., Rao, R., Pirolli, P.: A Focus+Context Technique Based on Hyperbolic Geometry for Visualizing Large Hierarchies. In: Proc. of the ACM Conference on Human Factors in Computing Systems, pp. 401–408. Addison-Wesley, Reading (1995)
4. Yanai, K.: Mining Visual Knowledge on the World Wide Web for G eneric Image Classification. Trans. JSAI 19(5), 429–439 (2004) (in Japanese)
5. Iyama, A., Sunayama, W., Yachida, M.: Topic Collection Support by Clustering Web Pages based on Topical Independence. Trans. JSAI 19(6), 561–570 (2004) (in Japanese)
6. Crawford, E., Kay, J., McCreath, E.: Automatic Induction of Rule for e-mail Classification. In: Proc. of the 6th Australasian Document Computing Symposium (2001)
7. Ueda, Y., Narita, H., Kato, N., Hayashi, K., Nambo, H., Kimura, H.: An Automatic Email Distribution by Using Text Mining and Reinforcement Learning. Trans. IEICE Inf. & Syst. J87-D1(10), 887–898 (2004) (in Japanese)
8. Balter, O., Sidner, C.L.: Bifrost Inbox Organizer: Giving users control over the inbox. In: Proc. of the Second Nordic Congerence on Human-Computer Interaction (2002)

9. Matsuo, Y., Tomobe, H., Hasida, K., Nakashima, H., Ishizuka, M.: Social Network Extraction from the Web information. Trans. JSAI 20(1), 46–56 (2005) (in Japanese)
10. Sano, K., Sayama, H.: BisNet:An Information Sharing System Using Bookmarks of Web Browsers. Trans. JSAI 20(4), 281–288 (2005)
11. Gmail, `http://mail.google.com/mail/`
12. Takaki, T., Fujii, A., Ishikawa, T.: Associative Document Retrieval by Query Subtopic Analysis and its Application to Patent Search. IPSJ Journal 46(4), 1074–1081 (2005)
13. Toyoda, M., Yoshida, S., Kitsuregawa, M.: Web Community Chart: A Tool for Navigating Numerous Web Pages by Related Topics. Trans. IEICE Inf. & Syst. J87-D1(2), 256–265 (2004)

Face Image Annotation in Impressive Words by Integrating Latent Semantic Spaces and Rules

Hideaki Ito, Yuji Kawai, and Hiroyasu Koshimizu

School of Information Science and Technology, Chukyo University
101 Tokodachi, Kaizu-cho, Toyota, Aichi, 470-0393 Japan
{itoh@sist,h10806m@st,hiroyasu@sist}.chukyo-u.ac.jp

Abstract. This paper describes a mechanism to annotate face images in impressive words which express their visual impressions. An annotation mechanism is developed by integrating latent semantic indexing, decision trees, and association rules. Moreover, visual and symbolic features of faces are integrated, which are corresponding to lengths and/or widths of face parts and impressive words, respectively. Relationships among these features are represented in a latent semantic space, their direct relationships in decision trees, and co-occurrence relationships among symbolic features in association rules, respectively. Efficiency of annotation results is improved by integrating these mechanisms, since their features are utilized effectively.

Keywords: face image annotation, impression, latent semantic indexing, latent semantic space, decision tree, association rule.

1 Introduction

In recent, several types of face image processing systems are developed, such as face recognition systems, multimodal interfaces, etc. In especial, to develop a face image database is required according to the progress of such processing systems. Then, face images are retrieved not only in terms of visual features of face images, but also in terms of words which represent visual impressions of face images. Therefore, it is necessary that suitable words are assigned to face images, i.e., face images are annotated in impressive words.

A face image annotation system is being developed, named FIARS (Face Image Annotation and Retrieval System)[5]. Its annotation mechanism is realized by integrating three mechanisms based on latent semantic spaces constructed by latent semantic indexing[6], association rules and decision trees[8]. So far, each mechanism is developed independently, and these mechanisms were not yet integrated[5]. The latent semantic space consists of descriptions of face images and words. Co-occurrence relationships among words are specified in association rules. These rules are useful for inferring additional words. Decision trees specify requirements on visual features to assign a specific word. By integrating these mechanisms systematically, annotation results are able to be improved since these mechanisms work complementarily.

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 591–598, 2009.

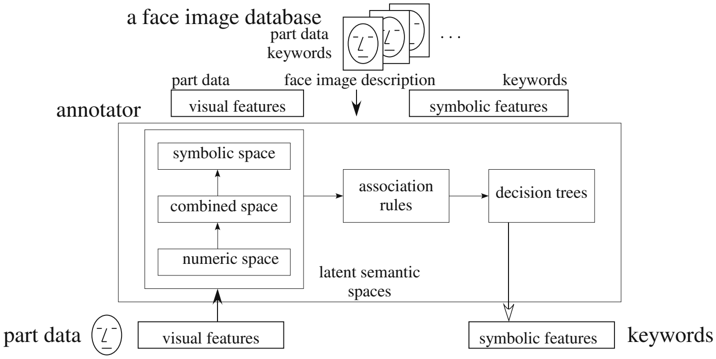


Fig. 1. An overview of an annotation mechanism and components of FIARS

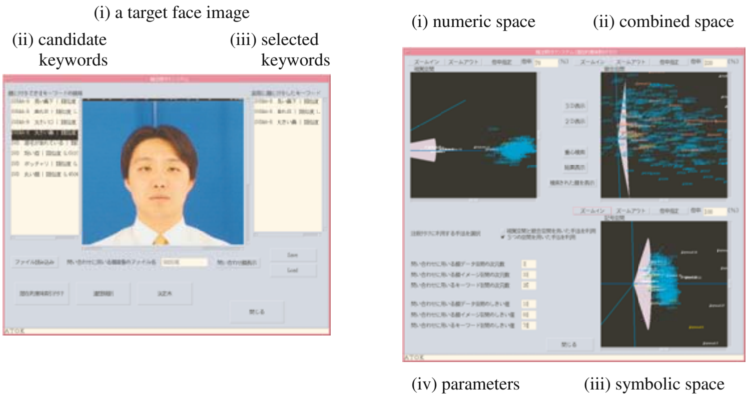
In recent, many mechanisms for retrieving and for annotating face images are developed[1]. In order to retrieve face images, to annotate them in keywords is required, like usual natural images[2]. Person identification is to identify the name of a person. Emotions indicated in faces are analyzed based on the facial action coding system[7,3]. [3] utilizes latent semantic analysis. Moreover, a face retrieval mechanism using mental words is developed[4].

This paper is organized as follows. Section 2 shows an overview of the system. Section 3 presents an annotation mechanism. Experimental results are shown in Sec. 4. Finally, concluding remarks are described in Sec. 5.

2 An Overview of FIARS

Face images are described in impressive words which depict visual impressions inspired from faces and their face parts. In current, a sort of impressive words is restricted. They express sizes, lengths or shapes of face parts, e.g., a round face, a thin lip, etc. On the other hand, lengths and/or widths of face parts are measured, they are called part data. Annotation is to assign and to find impressive words as keywords based on part data.

Figure 1 shows an overview of FIARS. This system consists of three mechanisms for annotation, and a face image database. These annotation mechanisms are developed based on latent semantic spaces, association rules and decision trees. They are constructed from the face image database. The latent semantic spaces consist of three spaces which are a numeric space, a combined space and a symbolic space. The numeric space is constructed from part data only, the combined space from part data and keywords, and the symbolic space from keywords, respectively. Moreover, association rules specify co-occurrence relationships among keywords. When association rules are applied to a set of existing keywords, some additional keywords are obtained. Furthermore, decision trees are constructed from both part data and keywords. Decision trees represent conditions on part data, which specify whether a keyword is able to be assigned to a face image, or not. On the other hand, the previous system shown in [5] utilizes



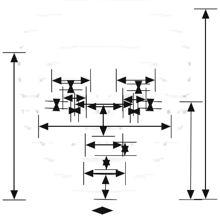
(a) An example of a window for annotating a face. (b) An example of windows for showing latent semantic spaces.

Fig. 2. Examples of screens for annotating a face image and for showing spaces

only a numeric space and a combined space. The symbolic space is constructed for making clear interrelationship among keywords, and for trying to improve efficiency of retrieved keywords. Moreover, latent semantic spaces, association rules and decision trees function independently each other in [5]. An entire annotation process has to be controlled by an individual user suitably. However, in this system, the system controls the procedure for applying each mechanism. They are working cooperatively, see Sec. 3.

This system provides some windows as an interface for assisting in annotation. Figure 2 (a) shows the main window of this system. (i) shows the face image to be annotated. (ii) and (iii) show candidate keywords inferred by the system, and selected keywords by a user, respectively. On the other hand, (i), (ii) and (iii) in Fig. 2(b) show a numeric space, a combined space and a symbolic space, respectively. (iv) shows some parameters for achieving keyword assignment.

Latent semantic spaces, association rules and decision trees are constructed from a face image database. Figure 3 (a) shows 24 measured places as part data. (b) shows an example of part data. Face image I_d is represented in terms of a vector, which consists of two vectors $(\boldsymbol{v}_d; \boldsymbol{w}_d)$. They are called a part vector and a keyword vector, respectively. A part vector is $\boldsymbol{v}_d = (v_{d,1}, \ldots, v_{d,24})^T$. For constructing a combined latent semantic space, a normalized part vector is computed from a part vector, $\boldsymbol{v}'_d = (v'_{d,1}, \ldots, v'_{d,24})^T$, where $v'_{d,j}$ is a normalized value of $v_{d,j}$. This value is computed as $v'_{d,j} = (v_{d,j} - \mu_j)/\sigma_j + 1/2$, where μ_j and σ_j are the mean value and the standard derivation of face part j. On the other hand, a keyword vector is $\boldsymbol{w}_d = (w_{d,1}, \ldots, w_{d,43})^T$. Each element $w_{d,j}$ is 1 or 0. They represent whether keyword j is assigned to face image I_d, or not, respectively. Furthermore, when a decision tree is constructed, discretization is applied to part data. Part datum $v_{d,j}$ is normalized as $v''_{d,j} = (v_{d,j} - \mu_j)/\sigma_j$. This value is transformed into one symbolic value, a, b or c. They are interpreted as small/short, normal and large/long, respectively. Let α be a threshold. If $v''_{d,j} < -\alpha$ then the value is a, $-\alpha \leq v''_{d,j} \leq +\alpha$ then b, and $+\alpha < v''_{d,j}$ then c.



length of the pupil of the left eye	1.2
length of the pupil of the right eye	1.3
length between the two pupils	6.0
length between the eyes	3.5
length of the left eye	3.2
. . .	
width of a face	13.2
length of a face	24.6
length of a face in visible	17.8
width of the chin	3.5

(a) part data.

(b) part data, lengths/widths of 24 places of a face.

Fig. 3. Part data and examples

If this threshold is 0.38, both a and c are assigned to about 35% of face images, respectively. The rest are assigned b.

3 An Annotation Mechanism

Figure 4 shows an overview of an annotation procedure. At first, a target face image to be annotated is given, which is specified in its part data.

Next, some keywords are obtained using three latent semantic spaces. To achieve this, the dimensions of spaces and the thresholds (see θ_1, θ_2 and θ_3 in Fig. 4) for query processing are specified. A part vector of a given face image is treated as a query vector(f_t). Some similar face images to a target are obtained using the numeric space. The vectors of the similar face images are found in the combined space. The centroid vector of them(f_c) is computed, which is used as a query vector for seeking keywords in the combined space. After query processing, some keywords are retrieved. The keyword vectors of obtained keywords are found in the symbolic space, the centroid vector of them(k_c) is computed, like the above described procedure. As this result, a retrieval result is obtained, which is a set of pairs of keyword k_i and its weight w_i, $K = \{< k_1, w_1 >, \ldots, < k_n, w_n >\}$. The weight is similarity weight between the keywords and the given face image. This similarity weight is computed by a cosine similarity measure.

To expand K, association rules are applied to it. An association rule is represented in $A_i : t_{i,1}, \ldots, t_{i,m} \rightarrow t_{i,m+1}, \cdots, t_{i,m+l}, (Sup_i, Con_i)$, where $t_{i,j}$, Sup_i and Con_i are keywords, support and confidence. Confidence is interpreted as a certainty factor. Rule A_i is applied to K, when all keywords appeared in the left-hand side of A_i are members of K. If so, words $t_{i,m+1}, \cdots, t_{i,m+l}$ are added to K. The weights of added keywords are $max\{w_{i,1}, \ldots, w_{i,m}\} * Con_i$, where $w_{i,1}, \ldots, w_{i,m}$ are weights of $t_{i,1}, \ldots, t_{i,m}$.

Finally, decision trees are applied to individual members of K. Rule D_i obtained based on a decision tree is represented as $D_i : t_i \leftarrow p_{i,1}, \cdots, p_{i,m}, ER_i$, where t_i, $p_{i,j}$ and ER_i are words, a condition element related to a face part and an error ratio of the rule. Condition-part is constructed based on decision trees using a specified error ratio[8]. After this, an actual error ratio of each obtained rule, ER_i is computed. Then, $1 - ER_i$ seems its certain factor CF_i. For k_i in K, it is checked whether a given face image satisfies the condition-part of rule D_i

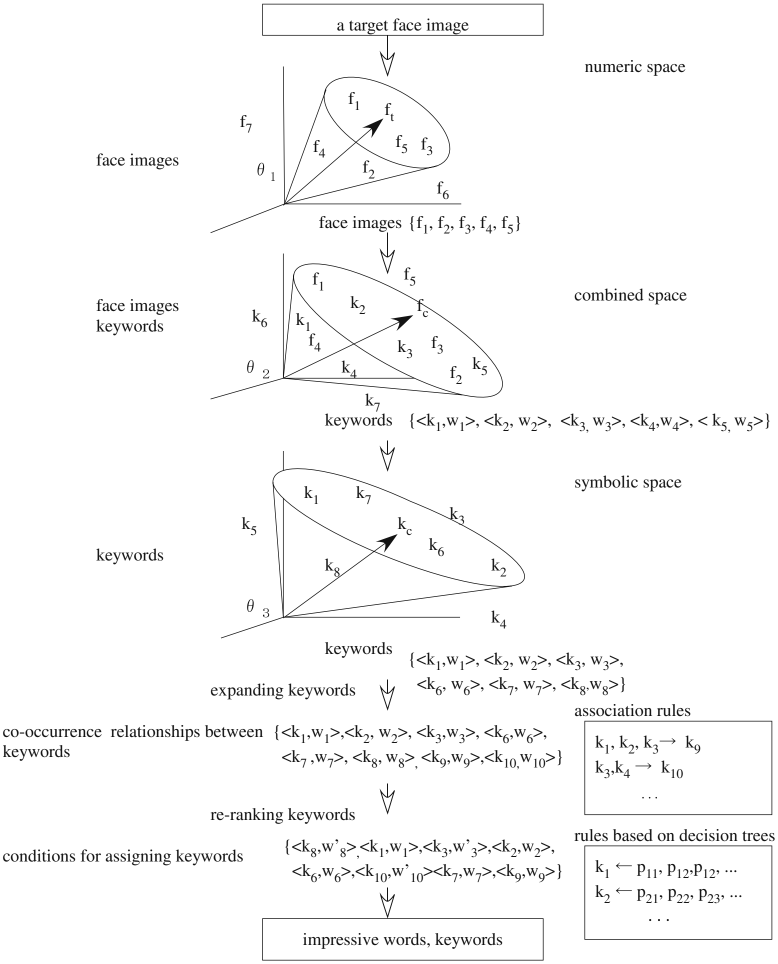

a target face image
numeric space
face images
f_7
f_1
f_t
f_4
f_5 f_3
θ_1
f_2
f_6
face images $\{f_1, f_2, f_3, f_4, f_5\}$
combined space
face images
keywords
k_6
f_1
f_5
k_2
f_c
k_1
f_4
k_3
f_3
k_5
θ_2
k_4
f_2
k_7
keywords $\{<k_1,w_1>, <k_2, w_2>, <k_3, w_3>, <k_4,w_4>, <k_5, w_5>\}$
symbolic space
keywords
k_5
k_1
k_7
k_3
k_c
k_6
k_8
k_2
θ_3
k_4
keywords $\{<k_1,w_1>, <k_2, w_2>, <k_3, w_3>, <k_6, w_6>, <k_7, w_7>, <k_8,w_8>\}$
expanding keywords
association rules
co-occurrence relationships between keywords
$\{<k_1,w_1>,<k_2, w_2>, <k_3,w_3>, <k_6,w_6>, <k_7 ,w_7>, <k_8, w_8>, <k_9,w_9>,<k_{10},w_{10}>\}$
$k_1, k_2, k_3 \rightarrow k_9$
$k_3,k_4 \rightarrow k_{10}$
. . .
re-ranking keywords
rules based on decision trees
conditions for assigning keywords
$\{<k_8,w'_8>,<k_1,w_1>,<k_3,w'_3>,<k_2,w_2>, <k_6,w_6>,<k_{10},w'_{10}><k_7,w_7>,<k_9,w_9>\}$
$k_1 \leftarrow p_{11}, p_{12},p_{12}, \ldots$
$k_2 \leftarrow p_{21}, p_{22}, p_{23}, \ldots$
. . .
impressive words, keywords

Fig. 4. A conceptual overview of a procedure for assigning keywords

for k_i, or not. When its conditions are satisfied with the given face image, the weight of k_i is replaced to CF_i if its weight w_i is lower than CF_i.

Now, an example of the process to assign keywords is described. Let a face image be given as a target, shown in Fig. 2(a)(i). Also, some parameters are specified, see Sec. 4. At first, some keywords are captured by using three latent semantic spaces. Four keywords are obtained as follows; (1) 'dropping eyebrow', (2) 'short eyebrow', (3) 'chubby face' and (4) 'round face', in the descending order. Although weights of these keywords are computed, they are not shown here for simplicity. Next, by applying the association rules to the obtained keywords, other four keywords are derived. They are (5) 'long length between the nose and the upper lip', (6) 'large nose', (7) 'dropping eye' and (8) 'large mouse' as an ordered list. Many of these keywords seem suitable to the given face image. These two kinds of keywords are arranged by their weights, the ordered keywords are

(5), (6), (1), (7), (2), (8), (3) and (4). Continuously, decision rules are applied to each keyword for confirming individual keywords. Finally, the ordered list is obtained, which is {(5), (7), (8), (6), (1), (2), (3), (4)}. This list is presented to a user as candidate keywords. The keywords obtained using association rules are located higher in this case. Moreover, it is considered that five keywords among retrieved eight keywords are suitable, which are (5), (7), (6), (3) and (4). Figure 2 shows two windows used in this process.

4 Experimental Results

Five types of experiments are tried as follows:

(1) LSS. Three latent semantic spaces are used, only.
(2) LSS, association rules. After keywords are obtained using (1), association rules are applies to them.
(3) decision tree, association rule. After keywords are obtained by applying decision trees, association rules are applied to them.
(4) decision tree. Keywords are obtained by applying decision trees, only.
(5) top 9. After keywords are retrieved by (2), decision rules are applied. The nine keywords in a higher rank among the retrieved keywords are tested, i.e., keywords are obtained by the proposed method in Sec. 3. Moreover, one existing face image description defined in the face image database has 8.7 keywords on average. So, nine keywords are evaluated.

Some parameters are required for building latent semantic spaces, association rules and decision trees, and for achieving keyword retrieval, as shown in Table 1. The dimensions and the threshold for the numeric space are 3 and 10°, respectively. A cumulative contribution ratio is used for deciding dimensions of a combined space and a symbolic space. When the cumulative contribution ratio is over 0.8, the number of cumulated singular values is treated as the dimensions of these spaces. On the other hand, the minimum support and the minimum confidence are specified for association rules. Moreover, the threshold used in discretization and the error ratio are specified for decision trees.

In each experiment, 30 face images are given as targets. Retrieval keywords are tested in sense of precision and recall, which are defined as *precision = the number of retrieved correct keywords / the number of retrieved keywords* and

Table 1. Some parameters for setting FAIRS

latent semantic space	dimension	threshold(a degree)
numeric space	3	10
combined space	33	80
symbolic space	26	70
association rule	minimum support	minimum confidence
	0.1	0.4
decision tree	threshold for discretization	error ratio
	0.38	0.2

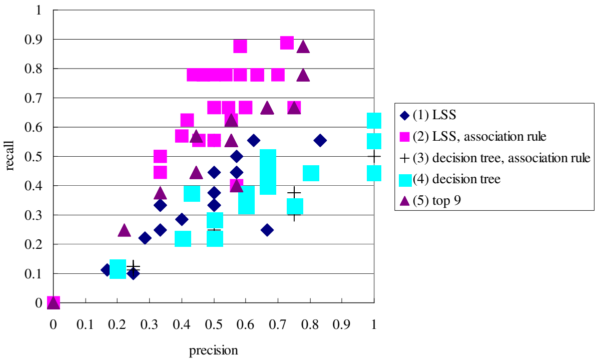


Fig. 5. Precision and recall of retrieval keywords

Table 2. Mean values of precisions and recalls of retrieval keywords

	precision	recall
(1) LSS	0.52	0.36
(2) LSS, association rule	0.53	0.68
(3) decision tree, association rule	0.63	0.35
(4) decision tree	0.69	0.31
(5) top 9	0.57	0.57

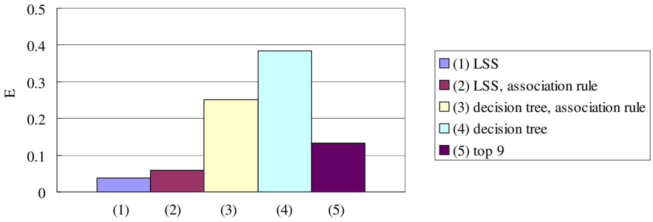


Fig. 6. Accuracy of retrieval keywords

recall = the number of retrieved correct keywords / the number of keywords assigned to a target in advance. Precisions and recalls in each experiment are shown in Fig. 5, and their mean values are summarized in Table 2. Recall in (2) is better than one in (1) by applying association rules. Precisions in (3) and (4) are better than ones in (1) and (2). The result using the decision trees is more precise than others. The result using (2) is obtained using the latent semantic spaces and association rules, however, decision trees are not applied to the retrieved keywords since their order is not evaluated in this case.

To evaluate ranked retrieval keywords, accuracy is computed. Accuracy is defined as $E =$ *(the number of correct keywords − the number of wrong keywords) / the number of retrieved keywords)*$(-1 \leq E \leq 1)$[9]. The result in each experiment is shown in Fig. 6. Recall in (2) is better than others, so, recalls in (3) and (4) are worse than one in (2), as shown in Table 2. However, accuracies in (3) and (4) are better than ones in (1) and (2). As shown in this table and Fig. 6, balance between precision and recall in (5) is better than others. Therefore, the proposed method is more effective than the method which utilizes only either latent semantic spaces or decision trees.

5 Concluding Remarks

This paper describes a mechanism for annotating face images in impressive words, which is developed by integrating latent semantic spaces, association rules and decision trees. Precision and recall of final retrieval keywords are more effective than ones using a single method. On the other hand, to develop some mechanisms is planed for improving capability of the system. It is required that some parameters are (semi-)automatically determined. Moreover, to develop a mechanism for retrieving appropriate face images using impressive words is necessary. Furthermore, the interface makes easy to understand assignment process.

Acknowledgement

The face images are used by permission of Softopia Japan, Research and Development Division, HOIP Laboratory.

References

1. Chellappa, R., Wilson, C.L., Sirohey, S.: Human and Machine Recognition of Faces: A Survey. Proceedings of the IEEE 83(5) (1995)
2. Datta, R., Joshi, D., Li, A., Wang, J.Z.: Image Retrieval: Ideas, Influence, and Trends of the New Age. ACM Computing Survey 40(2) (2008)
3. Fasel, B., Monay, F., Gatia-Perez, D.: Latent Semantic Analysis of Facial Action Code for Automatic Facial Expression Recognition. In: Proc. MIR. ACM, New York (2004)
4. Fang, Y., Geman, D., Boujemaa, N.: An Interactive System for Mental Face Retrieval. In: Proc. MIR. ACM, New York (2005)
5. Ito, H., Kawai, Y., Koshimizu, H.: Face Image Annotation based on Latent Semantic Space and Rules. In: Lovrek, I., Howlett, R.J., Jain, L.C. (eds.) KES 2008, Part II. LNCS, vol. 5178, pp. 766–773. Springer, Heidelberg (2008)
6. Manning, C.P., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press, Cambridge (2008)
7. Pantic, M., Rothkrantz, L.J.M.: Facial Action Recognition for Facial Expression Analysis from Static Face Images. IEEE Tran. on SMC. Part B 34(3) (2004)
8. Tan, P.-N., Steinback, M., Kumar, V.: Introduction to Data Mining. Addison-Wesley, Reading (2005)
9. Wang, X.-J., Zhang, L., Ma, W.-Y.: Annotating Images by Mining Image Search Results. IEEE Tran. on PAMI 30(11), 1919–1932 (2008)
10. Zhang, Z., Zhang, R.: Multimedia Data Mining, A Systematic Introduction to Concepts and Theory. CRC Press, Boca Raton (2009)

Sketch Learning Environment for Human Body Figure by Imitative Drawing

Masato Soga, Takahisa Fukuda, and Hirokazu Taki

Faculty of Systems Engineering, Wakayama University
930 Sakaedani, Wakayama, 640-8510 Japan
soga@sys.wakayama-u.ac.jp

Abstract. We developed an interactive learning environment for imitative figure sketching. Figure sketching is more difficult for novice than other sketching. People can easily find errors of figure sketch, since human is sensitive to human body figure. There are some important points to draw figure sketch. In this paper, we focus on length and angle between junctions. After learners draw human body figure by imitative sketching, the learning environment diagnoses the lengths and angles between junctions of drawn figure sketch. The environment shows scores of the learners' figure sketch and some advice. We evaluated the environment with some learners.

Keywords: Learning environment, Sketch, Imitative drawing, Figure painting, Skill.

1 Introduction

There exist many systems and software that support drawing or painting on computers. For example, Baxter developed an excellent system called DAB that assists a user in painting on a virtual paper on a computer [1]. Although DAB is an excellent system, it cannot be used for learning support, since it does not have a function for diagnosing the sketches by users. Learning support for drawing or painting is a task that differs from drawing support or painting support. Functions for diagnosis and advice are required for learning support. Our project was the first learning environment that could diagnose a learner's sketch and then provide advice.

We have previously developed various sketch learning support environments. [2-6] are the environments that use pre-defined motifs. Therefore, the learning environment is designed motif-dependently. The motifs are dish and glass. [7] is a learning environment that trains perspective. It is a motif-independent learning environment.

Although we developed various learning environment, there was no learning environment that trains learners to draw human figure sketch. Therefore, we developed a learning environment that trains learners to draw human figure sketch by imitative drawing. The target is novice learners for learning human body figure sketch. The novice learners could be students who studies arts, and also could be students who don't study arts.

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 599–606, 2009.

2 Cognitive Discussion for Figure Drawing

Human interactions with objects comprises three main stages, namely, recognition, selection (or decision), and action (Fig. 1). This process is explained for the task of drawing a sketch as follows.

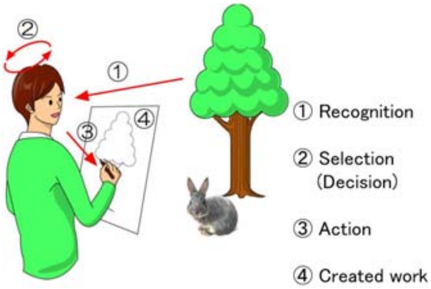


Fig. 1. Interaction between a learner and objects. The interaction comprises 3 stages, namely, (1) recognition, (2) selection (decision), and (3) action. A learner creates his/her work by repeating this process many times.

In the recognition stage, a learner perceives objects as a motif of drawing, and recognizes them. If the learner recognizes a tree, then he/she moves into selection (or decision) stage as follows. The learner thinks how to draw the tree. Then he/she decides to draw the edge of the tree first. He/she also decides to draw the shaft of the tree after drawing the edge. Thus, in the selection (or decision) stage, the learner selects (or decides) an appropriate action that he/she will acts in the next stage. Finally, in the action stage, the learner acts in accordance with the action that he/she selects (or decides) in the selection (or decision) stage.

If the motif is still objects like tree or dish, recognition of the motif is quite easy. However, if the motif is a human body, more precise recognition is required, since people can recognize the human body figure more sensitively than other still objects. Therefore, if a sketch of human body figure includes a small error, people easily find the error and feel the error larger than it really is.

To minimize the errors of human body figure sketch, the recognition stage is important for learners, since an error in the action stage often comes from an error in the recognition stage. Learners are required to recognize the human body figure more precisely to draw it without errors. A human body figure depends on a human pose. One of the most important matters to recognize a human pose is to recognize the junctions and the skeleton in the human pose. If learners can recognize the junctions and the skeleton precisely, he/she can draw them precisely quite easily on a paper. After drawing the junctions and the skeleton, he/she can quite easily draw the contours of the human body figure. Therefore, we focused on the junctions and the skeleton, and developed learning environment for drawing human body figure.

3 Learning Environment Design

Figure 2 shows the workflow of the learning environment. The learning environment consists of a PC, a monitor and a tablet Intuos 2 made by WACOM. The tablet has a grip pen for a learner to put in position coordinate values on a paper on the tablet.

At the beginning, the environment shows an example of human body figure paintings on the monitor. Then, the learning environment requires the learner to recognize comparative positions between junctions and the angle between the bones. After recognition of the junctions and the skeleton, the learner draws the skeleton first, then draws contours of the human body figure on a paper on the tablet.

After completing drawing, next stage is self-diagnosis stage. The learner puts in the position coordinate values by putting grip pen on the junction positions in the sketch on the tablet. Three junctions are, for instance, shoulder, elbow and wrist. After putting in three junction positions, the learning environment calculates ratio of bone lengths between junctions and the angle of the middle junction between two bones. The learning environment diagnoses these values by comparing with correct values. Then, the learning environment shows advice on the monitor according to the results of the diagnosis. The learner modifies his/her sketch according to the advice, and tries diagnosis again. Repeating this process, the learner can get drawing skill of human body figure.

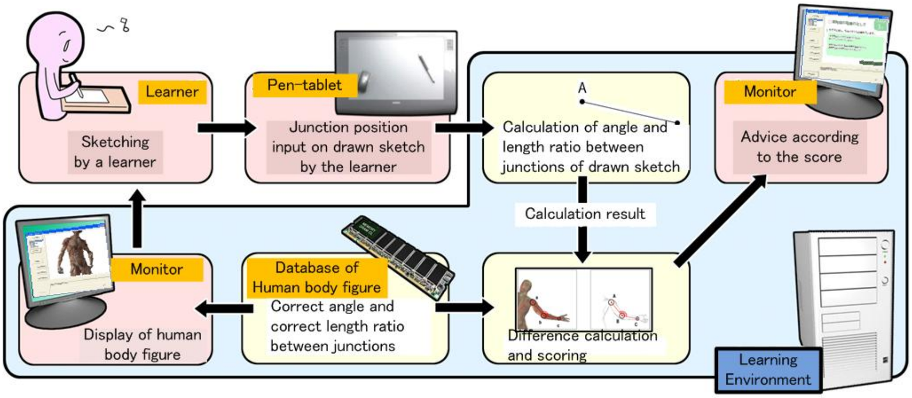
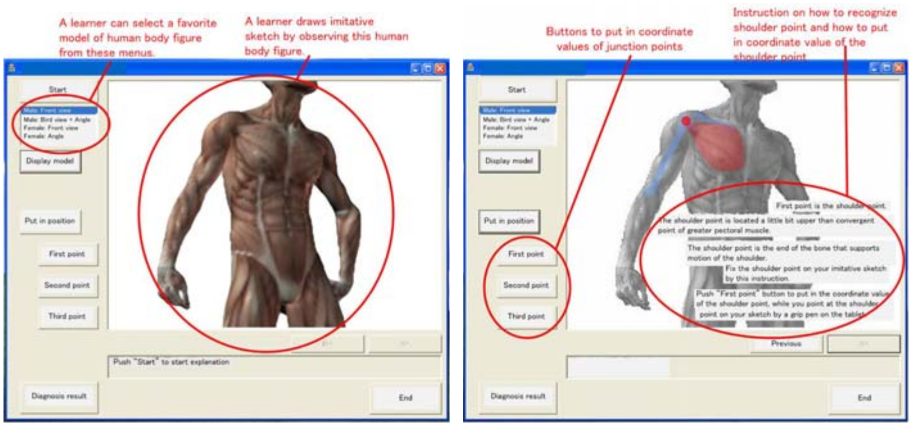

Learner
Sketching by a learner
Pen-tablet
Junction position input on drawn sketch by the learner
A
Calculation of angle and length ratio between junctions of drawn sketch
Monitor
Advice according to the score
Calculation result
Monitor
Display of human body figure
Database of Human body figure
Correct angle and correct length ratio between junctions
Difference calculation and scoring
Learning Environment

Fig. 2. Workflow of the learning environment

4 GUI of the Learning Environment

Figure 3 shows GUI of the learning environment. Actually, GUI of the original learning environment is made in Japanese. The GUI in figure 3 is indicated in English by synthesizing texts in English for explanation. Figure 3(a) shows a model of human body figure for imitative drawing. Figure 3(b) is a scene instructing the learner on how to put in each junction point coordinate value.

Actually, 2 sets of training and learning contents are prepared in the learning environment. One is male model sets and the other is female model sets. The learning environment can diagnose angle at elbow, and bone length ratio between shoulder, elbow and wrist for the male model set. On the other hand, the learning environment can diagnose angle at knee, and bone length ratio between hip, knee and ankle for the female model set.



A learner can select a favorite model of human body figure from these menus.

A learner draws imitative sketch by observing this human body figure.

Buttons to put in coordinate values of junction points

Instruction on how to recognize shoulder point and how to put in coordinate value of the shoulder point

Push "Start" to start explanation

First point is the shoulder point.

The shoulder point is located a little bit upper than convergent point of greater pectoral muscle.

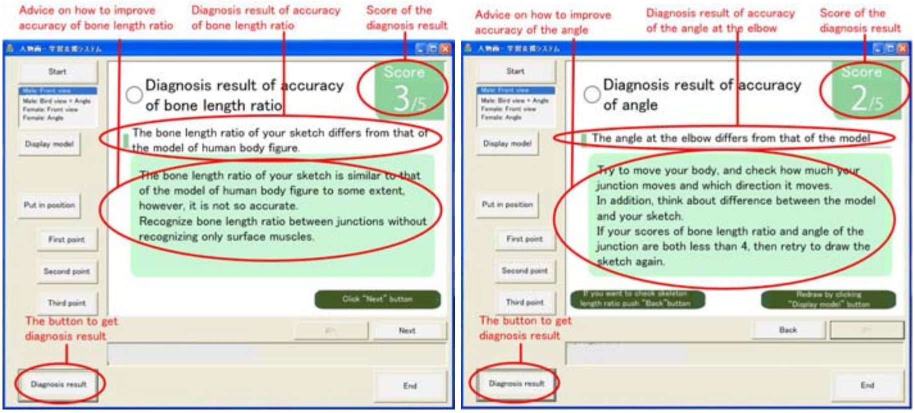
The shoulder point is the end of the bone that supports motion of the shoulder.

Fix the shoulder point on your imitative sketch by this instruction.

Push "First point" button to put in the coordinate value of the shoulder point, while you point at the shoulder point on your sketch by a grip pen on the tablet.

(a)A model of human body figure (b) Instruction on how to put in shoulder position

Fig. 3. GUI of the learning environment



Advice on how to improve accuracy of bone length ratio

Diagnosis result of accuracy of bone length ratio

Score of the diagnosis result

The button to get diagnosis result

Diagnosis result of accuracy of bone length ratio

The bone length ratio of your sketch differs from that of the model of human body figure.

The bone length ratio of your sketch is similar to that of the model of human body figure to some extent, however, it is not so accurate.
Recognize bone length ratio between junctions without recognizing only surface muscles.

Advice on how to improve accuracy of the angle

Diagnosis result of accuracy of the angle at the elbow

Score of the diagnosis result

The button to get diagnosis result

Diagnosis result of accuracy of angle

The angle at the elbow differs from that of the model

Try to move your body, and check how much your junction moves and which direction it moves.
In addition, think about difference between the model and your sketch.
If your scores of bone length ratio and angle of the junction are both less than 4, then retry to draw the sketch again.

If you want to check skeleton length ratio push "Back"button

Redraw by clicking "Display model" button

(a) Result of diagnosis of bone length (b) Result of diagnosis of angle

Fig. 4. Display of score and advice after diagnosis

Figure 4 shows examples of diagnosis result shown in English for explanation as well as figure 3. Figure 4(a) shows an example of diagnosis result of bone length ratio. The score indicates difference between drawn sketch and correct model. The score is indicated from 1/5 to 5/5. 5/5 is the best score, and it means there is almost no difference between drawn sketch and correct model.

5 Evaluation

We evaluated the learning environment by comparing experimental group with control group.

5.1 Goal of Evaluation Experiment

Goal of evaluation experiment is to confirm the learning effect by the learning environment compared with text instruction on printed papers. We tried to evaluate learning effect by not only novice but also experienced subjects.

5.2 Method of Evaluation Experiment

Figure 5 shows flow of evaluation experiment. 12 students in our university were subjects. 8 students of them were novice. The other 4 students were experienced subjects who had had experience of drawing sketches more than 30 hours totally. They were neither novice nor expert. 12 students were divided into two groups, experimental group and control group. Experimental group consists of 6 students. 4 students were novice, and 2 students were experienced subjects. Experimental group used the learning environment for training and learning. Control group also consists of 6 students. 4 students were novice, and 2 students were experienced subjects. Control group did not use the learning environment but read instruction on papers for training and learning.

Every student in every group had pre-test before training & learning, and had post-test after training & learning. Every student drew 3 different human body figures in training & learning. Models of human body figures in pre-test, post-test and training & learning are different. Every student answered questionnaire after post-test.

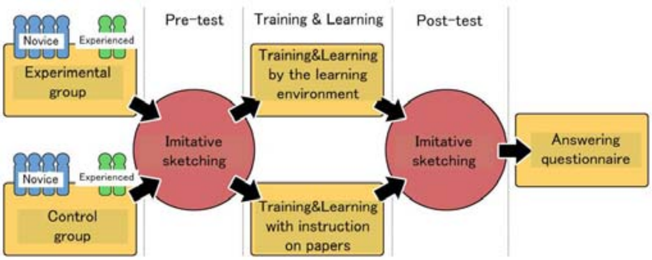


Fig. 5. Flow of evaluation experiment

5.3 Evaluation Method for Drawn Sketches

The learning environment has 2 sets of human body figure as training & learning contents. Figure 6 shows models of female human body figure with junctions and skeleton for evaluation and sketches by subject E in the experimental group (Most left: model for pre-test, Left: the sketch by subject E, Right: model for post-test, Most right: the sketch by subject E). Subject E had most learning effect in all subjects. Figure 7 shows models of male human body figure with junctions and skeleton for evaluation and sketches by subject I in the control group (Most left: model for pre-test, Left: sketch by subject I, Right: model for post-test, Most right: sketch by subject I) The sketch in post-test by subject I was worse than his sketch in pre-test in the evaluation by whole human figure. Skeletons and junctions in the figure 6 and 7 were not shown in the pre-tests and post-tests.

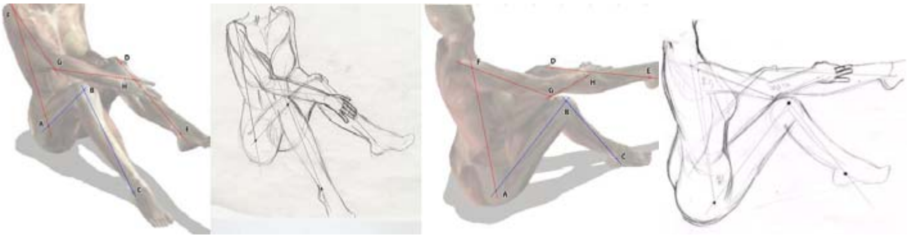


Fig. 6. Model of female human body figure with junctions and skeleton for evaluation and sketches by subject E in the experimental group (Most left: model for pre-test, Left: sketch by subject E, Right: model for post-test, Most right: sketch by subject E) Subject E had most learning effect in all subjects

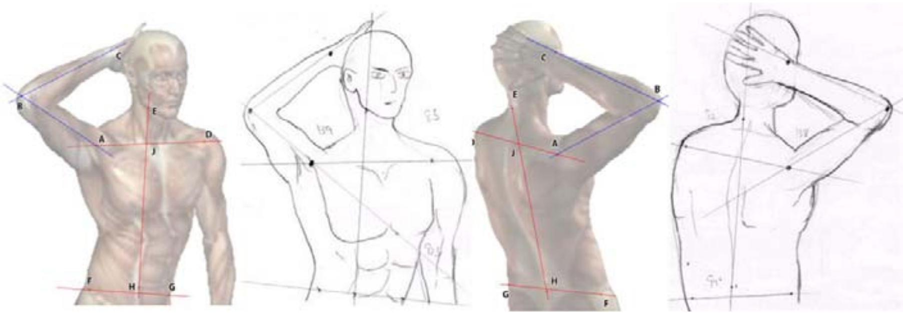


Fig. 7. Models of male human body figure with junctions and skeleton for evaluation and sketches by subject I in the control group (Most left: model for pre-test, Left: sketch by subject I, Right: model for post-test, Most right: sketch by subject I) The sketch in post-test by subject I was worse than his sketch in pre-test by evaluation of whole human figure

Table 1 shows results of skill learning effect in the experimental group about the part of human body figure where the learning environment was able to diagnose and advise the learner. Table 2 shows those in the control group.

Length ratio is the value calculated by BC/AB. Alphabets correspond those in figure 6 and 7. Score of length ratio means accuracy of the length ratio of subject's sketch to the model. Score of angle means accuracy of ∠ABC of subject's ketch to the model. If the scores of length ratio and angle are both 100, it means the skeleton and the angle of the part of the sketch correspond perfectly to those of the model.

Comparing table 1 and table 2, unfortunately learning effect cannot be proved. Each group had one subject who enhanced drawing skill in both length ratio and angle (Subject E and subject H).

We also compared the experimental group with the control group by other values to cover whole human body figure. Specifically, the other values are ∠BAJ, ∠EJD, ∠JHG, (AB/AD), (AD/FG), (FG/EH) for male figures, and ∠BAF, ∠AFG, ∠FGH, (AB/DE), (DE/AF), (AF/FG), (FG/GH) for female figures. Table 3 shows evaluation results of whole human body figure in the experimental group. Table 4 shows those in the control group. Scores in pre-tests and post-tests are averages of 6values for a male figure, and averages of 7 values for a female figure.

Table 1. Result of learning effect at the trained part of human figure in the experimental group

Subject ID	Model		Pre-test	Post-test	Skill
Subject A (novice)	Male figure	Score of length ratio	99.5	94.3	Down
		Score of angle	88.7	93.9	Up
Subject B (Experienced)	Male figure	Score of length ratio	92.9	90.5	Down
		Score of angle	96.2	84.8	Down
Subject C (Experienced)	Male figure	Score of length ratio	73.2	94.2	Up
		Score of angle	96.9	63.6	Down
Subject D (novice)	Female figure	Score of length ratio	99.2	96.5	Down
		Score of angle	87.6	94.0	Up
Subject E (novice)	Female figure	Score of length ratio	80.5	92.6	Up
		Score of angle	90.4	96.6	Up
Subject F (novice)	Female figure	Score of length ratio	87.8	92.1	Up
		Score of angle	87.6	74.2	Down

Table 2. Result of skill learning effect at the trained part of human body figure in the control group

Subject ID	Model		Pre-test	Post-test	Skill
Subject G (novice)	Male figure	Score of length ratio	85.0	95.4	Up
		Score of angle	99.5	96.6	Down
Subject H (novice)	Male figure	Score of length ratio	89.3	94.7	Up
		Score of angle	73.4	95.8	Up
Subject I (Experienced)	Male figure	Score of length ratio	84.9	97.1	Up
		Score of angle	95.3	79.7	Down
Subject J (novice)	Female figure	Score of length ratio	68.6	95.5	Up
		Score of angle	96.9	86.9	Down
Subject K (novice)	Female figure	Score of length ratio	92.6	98.7	Up
		Score of angle	94.6	94.0	Down
Subject L (Experienced)	Female figure	Score of length ratio	94.6	68.0	Down
		Score of angle	91.1	97.1	Up

Table 3. Evaluation result of whole human body figure in the experimental group

Subject ID	Model	Pre-test	Post-test	Skill
Subject A (novice)	Male figure	93.4	90.0	Down
Subject B (Experienced)	Male figure	90.0	92.6	Up
Subject C (Experienced)	Male figure	87.2	89.0	Up
Subject D (novice)	Female figure	91.1	92.7	Up
Subject E (novice)	Female figure	90.8	94.1	Up
Subject F (novice)	Female figure	87.1	93.4	Up

Table 4. Evaluation result of whole human body figure in the control group

Subject ID	Model	Pre-test	Post-test	Skill
Subject G (novice)	Male figure	95.0	88.8	Down
Subject H (novice)	Male figure	94.0	83.1	Down
Subject I (Experienced)	Male figure	92.6	86.2	Down
Subject J (novice)	Female figure	85.5	83.5	Down
Subject K (novice)	Female figure	95.2	81.8	Down
Subject L (Experienced)	Female figure	93.3	84.8	Down

Table 3 indicates that every drawn sketch in post-test is better than pre-test except subject A in experimental group. On the other hand, table 4 indicates every drawn sketch in post-test is worse than pre-test. From these results, the learning environment is effective to some extent for imitative sketching of human body figure.

Table 4 indicates that every sketch in post-test is worse than pre-test in control group. We think that the reason of this result comes from no review of drawing and fatigue. The experimental group also could be tired of drawing. However, review and learning effect by the environment exceeded fatigue in the experimental group.

6 Conclusion

In this paper, we described an interactive learning environment for imitative figure sketching. We focus on the length ratio and angle between junctions. After learners draw figure by imitative sketching, the learning environment diagnoses the length ratios and angles between junctions of learners' figure sketch. The environment shows scores of the learners' figure sketch and gives some advice.

We evaluated the environment with some learners. The learning environment is effective to some extent for imitative sketching of human body figure.

Acknowledgments. This research was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (B), 19300282, 2009. We are grateful to Posemaniacs (http://www.posemaniacs.com/) for offering human body figures.

References

1. Baxter, W., Scheib, V., Lin, C.M., Manocha, D.: DAB: Interactive Haptic Painting with 3D Virtual Brushes. In: Proc. of the 28th annual conference on Computer graphics and interactive techniques, pp. 461–468 (2001)
2. Iwaki, T., Tsuji, T., Maeno, H., Soga, M., Matsuda, N., Takagi, S., Taki, H., Yoshimoto, F.: A Sketch Learning Support System with Automatic Diagnosis and Advice. In: Int. Conf. of Computers in Education 2005 (ICCE 2005), pp. 977–979 (2005)
3. Takagi, S., Matsuda, N., Soga, M., Taki, H., Shima, T., Yoshimoto, F.: An educational tool for basic techniques in beginner's pencil drawing. In: Proc. of Computer Graphics International 2003, pp. 288–293 (2003)
4. Soga, M., Matsuda, N., Taki, H.: A Sketch Learning Support Environment that Gives Area-dependent Advice during Drawing the Sketch. Transactions of the Japanese Society for Artificial Intelligence, 96–104 (2008) (in Japanese)
5. Soga, M., Maeno, H., Koga, T., Wada, T., Matsuda, N., Takagi, S., Taki, H., Yoshimoto, F.: Development of the Sketch Skill Learning Support System with Real Time Diagnosis and Advice showing Learner's Arm Motion Animation. Transactions of Japanese Society for Information and Systems in Education 24(4), 311–322 (2008) (in Japanese)
6. Soga, M., Matsuda, N., Takagi, S., Taki, H., Yoshimoto, F.: Sketch Learning Environment based on Drawing Skill Analysis. In: Apolloni, B., Howlett, R.J., Jain, L. (eds.) KES 2007, Part III. LNCS, vol. 4694, pp. 1073–1080. Springer, Heidelberg (2007)
7. Shojiguchi, Y., Soga, M., Matsuda, N., Taki, H.: Interactive Learning Environment for Drawing Skill Based on Perspective. In: Lovrek, I., Howlett, R.J., Jain, L.C. (eds.) KES 2008, Part III. LNCS, vol. 5179, pp. 695–700. Springer, Heidelberg (2008)

Design and Implementation of an Optimal Radio Access Network Selection Algorithm Using Mutually Connected Neural Networks

Mikio Hasegawa[1,2], Taichi Takeda[1], and Hiroshi Harada[2]

[1] Tokyo University of Science, Tokyo 102-0073, Japan
[2] National Institute of Information and Communications Technology (NICT), Yokosuka 239-0847, Japan

Abstract. We propose a distributed and autonomous algorithm for radio resource usage optimization in heterogeneous wireless network environment. We introduce optimization dynamics of the mutually connected neural network to optimize average throughput per the terminals and the load balancing among the radio access networks (RANs). The proposed method does not require a server to collect whole information of the network and compute the optimal state of RAN selections for each terminal. We construct a mutually connected neural network by calculating the connection weights and the thresholds of the neural network to autonomously minimize the objective function. By numerical simulations, we show that the proposed algorithm improves both the total and the fairness of the throughput per terminal. Moreover, we implement the proposed algorithm on an experimental wireless network distributively, and verify that the terminals optimize RAN selection autonomously.

1 Introduction

Various radio access networks (RANs) have been developed and deployed. The cellular phone networks provide ubiquitous services and available almost everywhere, but their data communication bit rate is not very high and the cost is relatively expensive. On the other hand, the wireless LAN systems provide high-speed and low cost network access and possible to put access points freely, but it is available only in limited areas, since the coverage of one access point is small. Each RAN has different feature on the connectivity, the transmission speed, the cost per bit, and so on. Therefore, the best RANs for the users to connect to the network always change depending on their situations and available RANs.

Recently, many of the networks are replaced by IP based networks. The cost of the voice over IP communication is much lower than the traditional circuit switched telephone networks. Moreover, the IP enables to exchange various kinds of data, web pages, e-mails, voice, streaming video, etc. By the increase of the demands for the Internet access, the most of RANs provide the Internet connectivity with the global IP access. This means that those RANs are connected to the same core network, the Internet.

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 607–614, 2009.

Across those RANs connected to the Internet, vertical handover technologies enable to switch the on-going sessions on one of the RAN to other RAN without interruption of the session [1,2]. The mobile IP[3] enables to switch the IP address of a mobile terminal for the session, seamlessly. IEEE 802.21 [4] provides a common interface to the upper layer protocols to control different kinds of RAN interfaces. By using those technologies, it becomes possible to seamlessly handover the sessions among different kinds of RANs.

The best RANs for each user always change depending on the user's location and situation, the network traffic load, the available and required QoS and so on. By vertical handover among different RANs, it becomes possible to optimize the radio resource usage of the whole wireless network environment. The architecture to exchange information required for radio resource usage optimization has been already standardized as IEEE 1900.4 [5]. To find the best RAN for all users to optimize the radio resource usage according to such information becomes a combinatorial optimization problem. There are a lot of researches to improve various factors, such as user throughput, load balancing, user QoS optimization and so on. To improve those factors, some of those researches utilize mutually connected neural network [6] to solve those combinatorial optimization problems [7,8]. Since the mutually connected neural network solves the problem by its autonomous and distributed dynamics, there is no need to run the algorithm at a centralized server with a heavy computational cost, and also no need to collect all of information of large scale network to one centralized server.

In this paper, we evaluate the effectiveness of the neural network approach by computer simulations and real experiments on an experimental wireless network. We apply this optimization approach to load balancing of the traffic, which improves fairness for the users and RANs. For the implementation for real experiment, we use the Cognitive Wireless Cloud (CWC) system [10], and verify the capability of the proposed approach.

2 Load Balancing Based on Neural Network Dynamics

There are a lot of factors to be optimized for the RAN selection in heterogeneous wireless network environment. In IEEE 1900.4 [5], various kinds of information are defined to be exchanged between the network side and the terminal side to choose the best wireless links to be connected. In this paper, we examine the performance and effectiveness of the neural network based algorithm by optimizing the load balancing while keeping maximization of the throughput per user.

First, we define the available throughput per user by an equation. For simplifying the experiments in this paper, we assume that all the terminals are communicating by a best-effort type application, and capacity of each access point is shared equally among the terminals connected to the same access point. Under such an assumption, available throughput for each terminal can be approximately defined as $T_i(t) = C_{h_{\mathrm{link}}(i)}/N^{\mathrm{AP}}_{h_{\mathrm{link}}(i)}(t)$, where $N^{\mathrm{AP}}_j(t)$ is the number of terminals connected to the access point j, C_j is the total of the throughput which the access point j can provide, $h_{\mathrm{link}}(i)$ is the access point which the terminal i is currently connecting.

In order to optimize the fairness and the total of the throughput at the same time, we use the following objective function,

$$E_{\mathrm{OBJ}}(t) = \sum_{i=1}^{N_{\mathrm{m}}} \frac{1}{T_i(t)} = \sum_{i=1}^{N_{\mathrm{m}}} \frac{N^{\mathrm{AP}}_{h_{\mathrm{link}}(i)}(t)}{C_{h_{\mathrm{link}}(i)}}, \tag{1}$$

where N_{m} is the number of mobile terminals in the network. By minimizing this simple function, the fairness and the total of the throughput can be optimized at the same time. Minimization of the reciprocal of the throughput $T_i(t)$ means maximization of the throughput. Moreover, the value of $E_{\mathrm{OBJ}}(t)$ becomes smallest in the case that all $T_i(t)$ becomes equal, when the total $\sum_{i=1}^{N_{\mathrm{m}}} T_i(t)$ is fixed. The problem is formulated as a combinatorial optimization problem that finds an optimal state of RAN selection for each terminal.

In order to minimize this simple function autonomously without collecting whole information and computing everything at one server, we introduce optimization dynamics of the mutually connected neural network. It is well-known that the energy function of the mutually connected neural network,

$$E_{\mathrm{NN}}(t) = -\frac{1}{2} \sum_{i=1}^{N_{\mathrm{m}}} \sum_{j=1}^{N_{\mathrm{AP}}} \sum_{k=1}^{N_{\mathrm{m}}} \sum_{l=1}^{N_{\mathrm{AP}}} W_{ijkl} x_{ij}(t) x_{kl}(t) + \sum_{i=1}^{N_{\mathrm{m}}} \sum_{j=1}^{N_{\mathrm{AP}}} \theta_{ij} x_{ij}(t). \tag{2}$$

always decreased and converges to a state corresponding to a minimum of this energy function by a typical neuronal update , such as the following equation,

$$x_{ij}(t+1) = \begin{cases} 1 & \text{for } \sum_{k=1}^{N_{\mathrm{m}}} \sum_{l=1}^{N_{\mathrm{AP}}} W_{ijkl} x_{kl}(t) > \theta_{ij}, \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

where, $x_{ij}(t+1)$ is the output of the (i,j) th neuron at time t, W_{ijkl} is the connection weight between the (i,j) th and (k,l) th neurons, θ_{ij} is the threshold of the (i,j) th neuron, respectively. The conditions for this convergence are that the weights of the self feedback connections are 0, $w_{ijij} = 0$, that the weights of the connections between the same pairs of neurons are equal, $w_{ijkl} = w_{klij}$ and that each neurons should be updated asynchronously.

To apply this neural network to solution search in a combinatorial optimization problem, first we have to define the relation between each solution and the firing pattern of the neural network. Since the problem is to find the wireless links which should be selected, we relate the firing of the (i,j) th neuron with an establishment of the wireless link between the terminal i and the access point j as shown in Fig. 1.

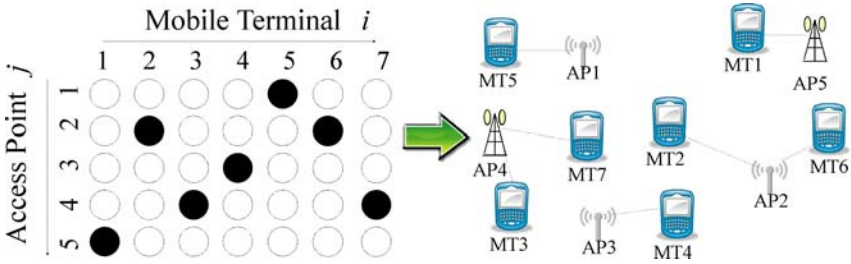


Fig. 1. Relation between firings of the neurons and establishments of the wireless links

Based on the relation described in Fig. 1, Eq. (1) can be transformed to a following form, a function of the state of neurons $x_{ij}(t)$.

$$E_{\mathrm{OBJ}}(t) = \sum_{i=1}^{N_{\mathrm{m}}} \sum_{j=1}^{N_{\mathrm{AP}}} \sum_{k=1}^{N_{\mathrm{m}}} \sum_{j=1}^{N_{\mathrm{AP}}} \frac{1}{C_j}(1-\delta_{ik})\delta_{jl} x_{ij}(t) x_{kl}(t) + \sum_{i=1}^{N_{\mathrm{m}}} \sum_{j=1}^{N_{\mathrm{AP}}} \frac{1}{C_j} x_{ij}(t). \quad (4)$$

By comparing Eqs. (2) and (4), we can obtain the connection weights and threshold to minimize Eq. (1), as follows,

$$W_{ijkl} = -2\frac{1}{C_j}(1-\delta_{ik})\delta_{jl}, \quad (5)$$

$$\theta_{ij} = \frac{1}{C_j}. \quad (6)$$

In the transformation from Eq. (1) to Eq. (4), we need to be careful to avoid self-feedback connection being larger than 0 to satisfy the condition of minimization on the energy function described above.

By autonomously updating each neuron by Eq. (3) with these obtained values, the state of the whole wireless network converges to an optimum state. In order to run this algorithm without centralized computation, we distribute the neurons to each corresponding terminal. Each terminal updates assigned neuron autonomously, makes a handover decision according to the state of the neurons, and hands over to the corresponding selected access point. This decentralized process optimizes the radio resource usage without any centralized computation.

In this paper's experiments, we assume that each terminal can establish only one wireless link with one access point at the same time. To satisfy such constraint, we need to control the number of firings one for each terminal. Usually, in the optimization neural network approach, we introduce a constraint term into the energy function. However, it sometimes could not be satisfied by local minimum problems and fatal infeasible solutions are obtained frequently. Therefore, in this paper, we introduce a maximum firing neuron,

$$x_{ij}(t+1) = \begin{cases} 1 & \text{if } y_{ij}(t+1) = \max\{y_{i1}(t+1), \ldots, y_{iN_{AP}}(t)\}, \\ 0 & \text{otherwise,} \end{cases} \quad (7)$$

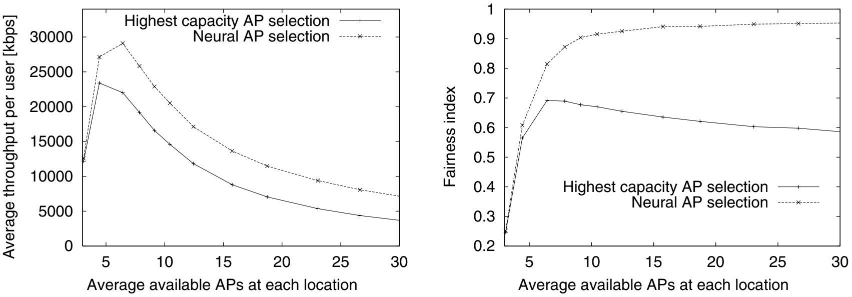


Fig. 2. Average throughput and fairness index of the proposed algorithm based on the neural network dynamics

where, $y_{ij}(t+1) = \sum_{k=1}^{N_{\mathrm{m}}} \sum_{l=1}^{N_{\mathrm{AP}}} W_{ijkl}^{A} x_{kl}(t) - \theta_{ij}^{A},$. By this update equation, we can keep one firing for each terminal.

The performance of the proposed optimization method for throughput maximization and load balancing is evaluated by comparing with the case that each terminal selects an access point which provides highest capacity in its location. To evaluate load balancing performance, we introduce the Jain's fairness index [11],

$$J = \frac{\left(\sum_{i=1}^{N_{\mathrm{m}}} T_i(t) \right)^2}{N_{\mathrm{m}} \sum_{i=1}^{N_{\mathrm{m}}} (T_i(t))^2}. \tag{8}$$

The results of the average throughput and the fairness index obtained by computer simulations are shown in Fig. 2. From the Fig. 2, we can confirm that the proposed method, which does not require any centralized computation, is effective for load balancing with improving the throughput, since both the average throughput and the fairness index could be improved. We have tested the proposed algorithm in the case with up to 200000 terminals.

3 Design and Implementation of a Neural Network Based RAN Selection Algorithm

In our implementation of the algorithm, the neurons are updated on each terminal distributively and autonomously. For each terminal, the neurons on the corresponding column in the left figure of Fig.1 are assigned. In real computation, each terminal has to calculate only for a limited number of neurons corresponding to available access points, which are detectable and reachable for the terminal, because we cannot establish a wireless link even if the neurons corresponding to unavailable links fire. By omitting those neurons corresponding to unavailable wireless links, scalability of the proposed system can be much improved.

In this paper, as an experimental wireless network to implement our algorithm, we use the CWC system [9,10]. This experimental system covers any functionality defined in IEEE 1900.4. In this system, various kinds of context information can be exchanged between the Network Reconfiguration Manager (NRM) on the network side and the Terminal Reconfiguration Manager (TRM) on the terminal side, via the Radio Enabler (RE).

In CWC system, three types of wireless network interfaces are defined. One of them is used as a common signaling channel [1,2] (RE in IEEE 1900.4) to exchange various context information. The second one is used to discover the available RANs. The last one is used as the RANs for the data communications. The terminal can seamlessly handover among different RANs, by switching care-of IP addresses for IP-in-IP capsuling packets between the mobility manager and the mobile terminal, that is almost the same procedure as the mobile IP[3]. The CWC system has additional function which enables IP address switching with multi-link aggregation support, but we do not consider it in this paper. We have developed the mobility manager, NRM, and the mobile terminals on the Linux operating system. To the mobile terminals installed on laptop PCs, we can attach various RAN interfaces which provides connectivity to the global Internet with a global IP address, by wireless LAN, 3G cellular systems, PHS and so on. In the following experiments, we use the wireless LANs for evaluating our algorithm based on the neural network dynamics.

In order to update each neuron, each terminal needs to obtain the state of other neurons which have non-zero connection weight with it. The states of those neurons can be derived from the information of each terminal's connecting access point, because the relation between the neurons' states and the selected wireless link is clearly defined so in Fig. 1. Therefore, in the implemented experimental system, each terminal receives such wireless link information of the other terminals, which have the neurons connected to the updating one. The information is transmitted to each mobile terminal's TRM via the NRM. Using such collected information, each terminal updates their neurons. According to the updated states of the neurons, each terminal autonomously selects an access point, and hands over to the selected one.

In the experiment described in the followings, we have used 4 wireless LAN access points and 8 mobile terminals. One of the access points and all of the terminals are placed in a same room. Other three access points are placed in another room. A scenario of this experiment is as follows. The mobile terminals, MT1, MT3, MT5, MT7 and MT8, are initially connected to one of those access points and communicating by the best-effort protocol. Between 50 to 100 seconds after the start of the experiment, the mobile terminals, MT2, MT4 and MT6 start their communications, and the additionally connected to the network. We observe the behavior of the algorithms in this scenario.

For a comparison, we have implemented and tested two algorithms. The first algorithm is that each terminal autonomously selects an access point which has

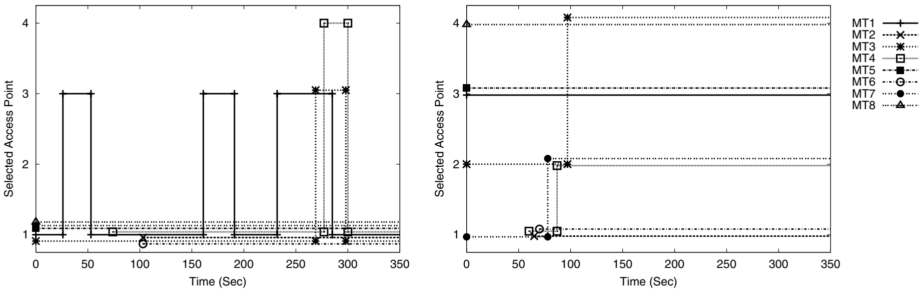


Fig. 3. Time series of selected access points by two algorithms, a generic algorithm in which each mobile terminal selects an access point with the strongest RSSI, on the left, and the proposed algorithm in which each mobile terminal selects an access point based on neural network dynamics optimizing the fairness, on the right

strongest RSSI. The second one is based on the proposed neural network dynamics described in Sec. II. Figure 3 shows the time series of the selected access points of each terminal in those two algorithms. From the Fig. 3 (left), a strongest RSSI selection tends to select the access point 1 which is located in the same room. On the other hand, in the case of proposed neural network based algorithm shown in Fig. 3 (right), selected access point is balanced. After joining of MT2, MT4 and MT6, it took only few neuron updates to converge to an optimal state that the loads of the access points are balanced.

It should be noted that our algorithm does not require any centralized computation to achieve the optimal state. Each terminal exchanges context information corresponding to the neuron state, updates the state of the neurons based on the context information, selects an access point according to the updated state of the neuron, and hands over to the selected the access point. This is distributed and autonomous process. Although we have tested this algorithm in the experimental network with a limited size, we have already shown that this algorithm performs well also in large-scale network by computer simulations.

4 Conclusion

In this paper, we have applied the distributed optimization dynamics of the mutually connected neural network to load balancing of the wireless networks. Since the proposed algorithm does not require any centralized computation, it is suitable for distributed networks, such as the heterogeneous wireless network environment in which each network is managed by different operator. We have shown that the proposed distributed algorithm can optimize fairness of the throughput in a large-scale wireless network, by computer simulations.

Furthermore, we have developed an experimental network to verify the effectiveness of the proposed optimization framework. By comparing our algorithm with a general network selection, we have shown that it is possible to optimize total network resource usage without any centralized decisions or computations.

In this paper, although we have applied our proposed framework only to the load balancing problem, it is applicable to various kinds of optimization problems. In Refs. [12] and [13], we have shown that our distributed algorithm based on the neural network can also optimize other objective functions, such as costs, power consumption and so on, by using more complicated model. In our future work, we are going to apply improved version suitable for more realistic cases in radio resource usage optimization in heterogeneous wireless networks, with evaluations in real experimental wireless networks.

References

1. Wu, G., Havinga, P., Mizuno, M.: MIRAI Architecture for Heterogeneous Networks. IEEE Comm. Mag., 126–134 (2002)
2. Inoue, M., Mahmud, K., Murakami, H., Hasegawa, M., Morikawa, H.: Novel Out-Of-Band Signaling for Seamless Interworking between Heterogeneous Networks. IEEE Wireless Commun. 11, 56–63 (2004)
3. Perkins, C.: IP Mobility Support for IPv4. IETF RFC 3344 (2002); Johnson, D., Perkins, C., Arkko, J.: Mobility Support in IPv6. IETF RFC 3775 (2004)
4. IEEE Std. 802.21: IEEE Standard for Local and metropolitan area networks- Part 21: Media Independent Handover (2009)
5. IEEE Std. 1900.4: IEEE Standard for Architectural Building Blocks Enabling Network-Device Distributed Decision Making for Optimized Radio Resource Usage in Heterogeneous Wireless Access Networks (2009)
6. Hopfield, J.J., Tank, D.W.: Neural Computation of Decisions in Optimization Problems. Biological Cybernetics 52, 141–152 (1985)
7. Gomez-Barquero, D., et al.: Hopfield Neural Network-Based Approach for Joint Radio Resource Allocation in Heterogeneous Wireless Networks. In: Proc. of IEEE Vehicular Technology Conference Fall (2006)
8. Garcia, N., Perez-Romero, J., Agutsi, R.: A new CRRM scheduling algorithm for heterogeneous networks using Hopfield Neural Networks. In: Proc. of WPMC (2006)
9. Harada, H., et al.: A Software Defined Cognitive Radio System: Cognitive Wireless Cloud. In: Proc. of IEEE Globecom (2007)
10. Ishizu, K., et al.: Design and Implementation of Cognitive Wireless Network based on IEEE P1900.4. In: Proc. of SDR workshop (2008)
11. Jain, R., Chiu, D., Hawe, W.: A Quantitative Measure of Fairness and Discrimination for Resource Allocation in Shared Computer System. DEC Technical Report 301 (1984)
12. Hasegawa, M., et al.: Application of Higher Order Neural Network Dynamics to Distributed Radio Resource Usage Optimization of Cognitive Wireless Networks. In: Köppen, M., Kasabov, N., Coghill, G. (eds.) ICONIP 2008, Part I. LNCS, vol. 5506, pp. 851–858. Springer, Heidelberg (2008)
13. Hasegawa, M., Tran, H., Miyamoto, G., Murata, Y., Harada, H., Kato, S.: User-Centric Optimum Radio Access Selection in Heterogeneous Wireless Networks based on Neural Network Dynamics. In: Proc. of IEEE Wireless Communication and Network Conference (2008)

Probabilistic Estimation of Travel Behaviors Using Zone Characteristics

Masatoshi Takamiya, Kosuke Yamamoto, and Toyohide Watanabe

Department of Systems and Social Informatics
Graduate School of Information Science, Nagoya University
Furo-cho, Chikusa-ku, Nagoya, 464-8603, Japan
{takamiya,yamamoto,watanabe}@watanabe.ss.is.nagoya-u.ac.jp

Abstract. There are many prior works of modeling travel behaviors. Most of them are investigated under the assumption that many kinds of data such as that of Person Trip (PT), which surveys travel behaviors, are available. Therefore, they do not consider an application to cities where the survey is not examined. In this paper, we propose a method for estimating travel behaviors using zone characteristics which is obtained from structural data of city. Focusing on dependent relationships between travel behaviors and city structure, we estimate the travel behaviors by means of the relationships. We first define trip and zone characteristics, and then introduce our method. With our method, we make use of Bayesian network constructed with PT data and the structural data. In addition, we show the effectiveness of our method through evaluation experiments.

Keywords: travel behaviors, Bayesian network, K2 algorithm.

1 Introduction

Generally, traffic simulations are utilized to measure the effect of new intelligent transportation systems in real world[1,2]. Although these studies exhibit several notable results, there are doubts about the validity of the results. Therefore, it is necessary to represent travel behaviors, which contain population flow, with computational simulations.

There are some studies to represent travel behaviors. Most of them estimate population flow using data of surveys which are examined on travel behaviors in real world, such as Person Trip (PT). Kitamura constructed a framework which simulates the living activities with activity-based approach which models individual travel behaviors[3]. Some studies adopt Neural network to the problem: for instance, Mozolin et al. compared[4] the performance of multilayer perceptron neural networks with that of maximum-likelihood doubly-constrained model, which is conventional model, for commuter travel behaviors, and Zhou et al. explored the application of back-propagation network to travel demand analysis [5]. However, those studies are under the assumption that many kinds of data such as residential information and survey data about travel behaviors

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 615–622, 2009.

of the city are available. It is expensive to examine the travel behaviors survey. Moreover, the survey is not examined in every metropolitan area. Therefore, in order to apply an estimation of general cities, we need a new method.

Our objective is to estimate travel behaviors using only structural data of city. The data contains position information about important facilities for travelers such as stations and schools. In addition, the data is able to be obtained without the survey about travel behaviors. We consider that there are general patterns of travel behaviors regardless of city and the patterns are involved by various factors such as time and location of the facilities. We estimate the behaviors using Bayesian network constructed by using the patterns.

The remainder of this paper is organized as follows. In Section 2, we refer to our approach. Section 3 mentions construction of Bayesian network. In Section 4, we explain how to estimate travel behaviors with our method. In Section 5, we report our experiments and results. Section 6 concludes this paper and offers our future work.

2 Approach

In order to achieve our objective, we propose a method for estimating travel behaviors using zone characteristics. Focusing on dependent relationships between travel behaviors and city structure, we extract and utilize trip patterns to estimate travel behaviors using zone characteristics. A city is divided into the several zones and the trip is defined as a personal movement from an origin zone to a destination zone for one purpose. Their characteristics are defined in Section 3.2. Using zone characteristics, we apply another city with only its structural data of city. With our method, probabilities which zones are selected as a destination zone with are calculated about an origin zone. Calculating them about overall origin zones, we can estimate travel behaviors in the entire city.

In order to represent the dependent relationship, we use Bayesian network. It is one of probabilistic models which represent conditional probability and indicate causal relationships by graph structure. Interpreting the dependent relationships as causal relationships, we are able to describe the travel behaviors using Bayesian network. Moreover, we discover the graph structure using K2 algorithm.

A flowchart of our method is shown in Fig.1. *Calculation* denotes a process which calculates probabilities for each zone characteristic, and *assignment* denotes a process which assigns the probabilities to all zones according to the characteristics. Bayesian network is constructed with PT data and structural data of city, and represents general travel behaviors. Therefore, we consider our method to be able to predict travel behaviors of other city.

3 Construction of Bayesian Network

Bayesian network has nodes corresponding to random variables, and represents static causal relationships among variables by the graph structure. For instance,

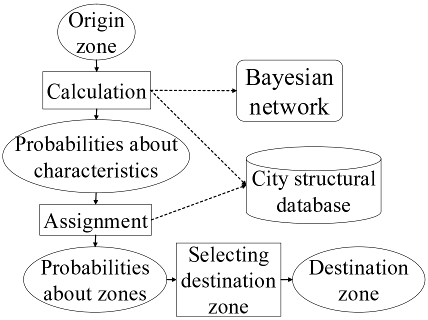


Fig. 1. Flowchart

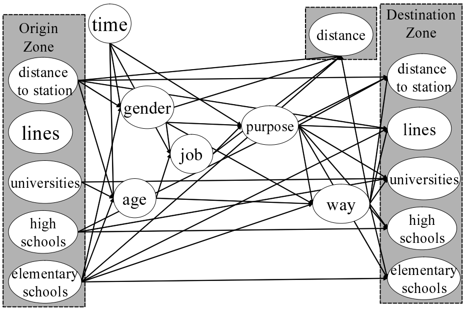


Fig. 2. Bayesian network

we assume that a node "a" corresponding to a variable "A" is conditioned by a node "b" corresponding to a variable "B". In this case, "b" is a parent node of "a" and there is a directed link from "b" to "a". Moreover, the each node has a Conditional Probability Table (CPT). CPT holds conditional probabilities for every combination of values its parent nodes have. Construction of Bayesian network has two phases. Firstly, entries in the CPT are calculated with research data. Secondly, an appropriate set of parent nodes is searched by algorithm for discovering graph structure.

3.1 Research Data

We utilize PT data of Chukyo metropolitan area where is in Japan as trip data, and structural data of Nagoya where is in the area as zone data. The PT examined a sample for 3% of the overall population in the area, which includes 259 zones in Nagoya, and was restricted to the residents of six or more ages. We extracted trips which move between zones in Nagoya from the PT data and the number of the trips is about 140000.

Bayesian network constructed with the data is shown in Fig. 2. *Lines* is a sum of train lines of stations in a zone, *distance* is spatial distance between centroid of an origin zone and that of destination zone, *time* is departure time. *Universities*, *high schools* and *elementary schools* are their sum.

3.2 Zone and Trip Characteristics

Zone. A zone is used as an origin or destination unit of a trip. This means that trips generating from inside a zone are aggregated as trips generating from a centroid of the zone. We divide a district into some zones according to PT. Zone characteristics represent an origin and destination zones in departure time of trips. The trips are dependent on the existences of important facilities for travelers. Therefore, we consider the characteristics to be represented by the information about the number and kinds of facilities in origin and destination zones for a trip. We define zone characteristics as follows:

- the number of elementary and junior high schools
- the number of high schools
- the number of universities
- sum of train lines of stations
- distance to the nearest zone, which has stations, from the zone (This is 0 if zone has station.)

Values of those parameters are obtained from the structural data.

Trip. We obtain trip patterns with PT data. We consider personal attributes and other attributes of trip. Therefore, we define trip characteristics as follows:

- age of travelers
- gender of travelers
- job of travelers
- travel purpose
- way to travel

where age, gender and job are personal attributes of trip. Values of those parameters are obtained from PT data.

3.3 Selecting Destination Order

We consider that there is dependent relationship among departure time, origin zone characteristics, personal attribute, travel purpose and destination zone characteristics. For instance, in the evening, students go home from school, and in the morning, students go to school and workers go to offices or shops. This denotes that personal attributes of trip, such as age and job, are dependent on the origin zone characteristics and its travel purpose is dependent on the personal attributes and origin zone characteristics. Therefore, we define destination selection order as follows:

1. origin zone characteristics
2. personal attribute
3. travel purpose
4. destination zone characteristics.

3.4 Discovery of Graph Structure

We utilize Bayesian network to represent dependent relationship between trip and zone. However, it is not clear to judge whether a node is linked to another node. We use K2 algorithm [6] which is a conventional algorithm for discovering graph structure. K2 algorithm searches an appropriate set of parent nodes for each node. The algorithm requires sets of candidate parent nodes for each node and tries to obtain optimal sets using greedy algorithm.

We use AIC as a scoring function for the search. AIC is a criterion for evaluating a model estimated by the maximum likelihood method, as follows:

$$AIC(M_k) = -2\sum_{n=1}^{N} \log p(x_n, \hat{\theta}_k) + 2p_k \tag{1}$$

where $\hat{\theta}_k$ denotes the parameter of model M_k, and p_k denotes the dimension number of $\hat{\theta}_k$ about a model set $M = \{M_1, M_2, ..., M_k\}$ and a data set $X = \{x_1, x_2, ..., x_N\}$. The value of AIC is small if the model estimates with high accuracy travel behaviors. The large dimension number fits the model the sample data too much. This problem is called overfitting. With the second term, AIC penalizes the dimension number and prevents the overfitting.

Generally, it is supposed that the nodes have a linear ordering to operate K2 algorithm. For each node, the set of its candidate parent nodes is built incrementally from the nodes which precede it in the linear ordering. Interpreting the selection of destination order as the linear ordering, we execute K2 algorithm.

4 Method for Estimation

Our method has three steps (Fig.1). Firstly, we calculate probabilities for each characteristic. Bayesian network is utilized to calculate them. Secondly, we assign the probabilities to all zones. Finally, we select a zone as destination from all zones according to the assigned probabilities for each zone.

In order to calculate the probabilities for each characteristic, we make use of Bayesian network shown in Fig.2. We calculate the probabilities using a production rule which is a fundamental rule of probability theory [7,8]. For instance, we assume that node "y" and node "z" are parent nodes of node "x". In this case, a probability of "x" is calculated recursively as follows:

$$p(x) = \sum_{z\in Z} p(z) \sum_{y\in Y} p(y)p(x|y,z) \tag{2}$$

Referring to CPT of node "x", we calculate $p(x|y,z)$.

In the second step, we assign the probabilities for each characteristic to all zones. A probability assigned to one zone is calculated as a joint probability of probabilities for each characteristic the zone has. Moreover, if some zones have the same characteristics, we assign the probability to the zones uniformly. For instance, if there are two zones which have the same number of schools and stations and the same distance from an origin zone, a half of selected probabilities are assigned to each zone.

5 Experiment

We show that travel behaviors are estimated with structural data of city using the proposed method. Therefore, we compare predicted performances in three

Table 1. Information which is available in three environments

	departure time	origin zone characteristics	personal attributes of trip characteristics	others of trip characteristics
Environment1	available	available	unavailable	unavailable
Environment2	available	available	available	unavailable
Environment3	available	available	available	available

environments where their available information is different respectively. We define two indices which are explained in Section 5.1 as the predicted performance.

Available information in the environments is shown in Table 1. Personal attributes denote *age*, *gender* and *job*. In addition, others denote *purpose* and *way*. If the information is available, observed and aggregated values are set in Bayesian network as evidence variables. Env.1, which is short for Environment1, simulates the target environment and Env.2 and Env.3 are ones for comparison with Env.1. The performance in Env.1 is essentially lower than those in others because the available data is restricted. The aim of this experiment is to show how the performance in Env.1 is close to those in others.

5.1 Evaluation Indices

Generally, several indices can be used to validate our model. We measure differences between observed and predicted distributions, and accuracy of predicted result. In order to measure them as the performance of the model, we use two following indices.

ARV. ARV is the average relative variance. This is normalized MSE (Mean Square Error), and often used to validate prediction models. In our study, we define it as follows:

$$\begin{aligned} MSE_t &= \frac{1}{|Z|^2} \sum_{o \in Z} \sum_{d \in Z} x(t,o)^2 (p(d|o,t) - \hat{p}(d|o,t))^2 \\ ARV_t &= \frac{MSE_t}{\sigma^2} \\ &= \frac{\sum_{o \in Z} \sum_{d \in Z} x(t,o)^2 (p(d|o,t) - \hat{p}(d|o,t))^2}{\sum_{o \in Z} \sum_{d \in Z} x(t,o)^2 (p(d|o,t) - \bar{p}(d))^2} \end{aligned} \quad (3)$$

where $x(t,o)$ denotes the number of observed trips from zone o at time t. $p(d|o,t)$ denotes an observed probability of trip toward zone d given all the trips from zone o at time t, and $\hat{p}(d|o,t)$ denotes a predicted probability. $\bar{p}(d)$ denotes an average probability of observed trip toward zone d through a whole day. ARV is 0 if the predicted distribution exactly equals to observed distribution, and is 1 if the model has standard performance. The standard performance means the performance of a model which always predicts average probability regardless of departure time and origin zone. The lower ARV is, the higher the predicted performance is.

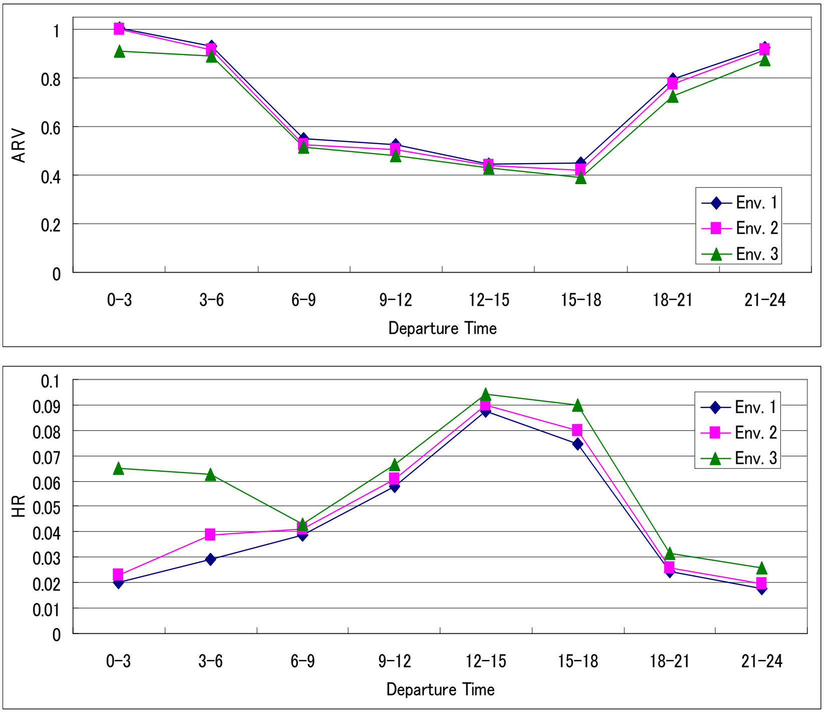


Fig. 3. Experimental results about ARV and HR

HR. HR is hit ratio. It is the ratio of the number of correctly predicted trips to the number of total trips. Unlike ARV, the higher HR is, the higher the predicted performance is.

5.2 Experimental Results

The results are presented in Fig.3. Upper graph shows the value of ARV, and lower graph shows the value of HR in tree environments through a whole day. Horizontal axis of the graph is departure time and vertical axis is value of the index at the departure time.

ARV in Env.1 is larger than that in Env.2 and HR in Env.1 is smaller than that in Env.2 at any time. Both these mean that the performance in Env.1 is less than that in Env.2. However, the difference between Env.1 and Env.2 is a few. In addition, the most difference is 0.04 about ARV and 0.01 about HR at 15-18. Available data in Env.1 is only structural data of city and those in Env.2 and Env.3 also include survey data of travel behaviors. Therefore, the result describes that travel behaviors are estimated efficiently only with the structural data.

However, the performance is remarkably low at 0-3, 3-6 and 21-24 when residents do not travel actively. We consider that this is because of the lack of sample data. The number of trips is few at the time. Therefore, we consider the sample data insufficiently. In addition, the performance is higher at 0-6 in Env.3 than those in Env.1 and Env.2. This means that it is difficult to predict trip purpose and way to travel especially at the time. This is because that at the time people behaves randomly in comparison with at other time. Moreover, HR is quit small in this experiment. This is because zone characteristics are not sufficient to select one from 259 zones. Therefore, we have to explore more adequate characteristics.

6 Conclusion

In this paper, we proposed a probabilistic method for estimating travel behaviors using zone characteristics. The method is based on a supposition that the travel behaviors are dependent on city structure. With the method, the dependent relationship is represented as Bayesian network. The experimental results show the effectiveness of our method in environments where structural data of city is only available. In our future work, we must explore more adequate characteristics and apply our Bayesian network to other city.

References

1. Uesugi, K., Mukai, N., Watanabe, T.: Optimization of Vehicle Assignment for Car Sharing System. In: Apolloni, B., Howlett, R.J., Jain, L. (eds.) KES 2007, Part II. LNCS (LNAI), vol. 4693, pp. 1105–1111. Springer, Heidelberg (2007)
2. Yamamoto, K., Uesugi, K., Watanabe, T.: Adaptive Routing of Cruising Taxis by Mutual Exchange of Pathways. In: Lovrek, I., Howlett, R.J., Jain, L.C. (eds.) KES 2008, Part II. LNCS (LNAI), vol. 5178, pp. 559–566. Springer, Heidelberg (2008)
3. Kitamura, R.: Applications of Models of Activity Behavior for Activity Based Demand Forecasting. In: Activity-Based Travel Forecasting Conference Proceedings (1997)
4. Mozolin, M., Thill, J.C., Usery, E.L.: Trip Distribution Forecasting with Multilayer Perceptron Neural Networks: A Critical Evaluation. Transportation Research Part B 34(1), 53–73 (2000)
5. Zhou, Q., Lu, H., Xu, W.: New Travel Demand Models with Back-Propagation Network. In: Proc. of ICNC 2007, vol. 3, pp. 311–317 (2007)
6. Cooper, G.F., Herskovits, E.: A Bayesian Method for the Induction of Probabilistic Networks from Data. Machine learning 9(4), 309–347 (1992)
7. Russell, S., Norvig, P.: Artificial Intelligence: A Modern Approach. Prentice-Hall, Englewood Cliffs (2002)
8. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, New York (2006)

A Web-Based Approach for Automatic Composition of an Insightful Slideshow for Personal Photographs

Kotaro Yatsugi[1], Naomi Fujimura[2], and Taketoshi Ushiama[2]

[1] School of Design, Kyushu University 4-9-1 Shiobaru, Minami-ku, Fukuoka, 815-8504 Japan
yatsugi@gospel.aid.design.kyushu-u.ac.jp

[2] Faculty of Design, Kyushu University 4-9-1 Shiobaru, Minami-ku, Fukuoka, 815-8504 Japan
{fujimura,ushiama}@design.kyushu-u.ac.jp

Abstract. Recently, the number of digital content objects is increasing rapidly with the progress of information technology. It has become important how we manage enormous digital content objects effectively and utilize them efficiently. Up to now, a lot of researches on digital content management have been reported. One of the important objectives of conventional management techniques is to search digital content objects that satisfy an information request of a user. This is based on the assumption that a user has one or more information requests. However, a user may have no information request explicitly when the user uses some kinds of devices for presenting digital content such as a digital photoframe. Such devices are expected to provide a presentation of digital content that rouses user's interest. In this paper, we introduce an approach for composing photo slideshow that attracts user's interest automatically.

1 Introduction

Digital cameras and cellular phones with camera have widely spread in human societies. In Japan, the number of households that have one or more digital cameras increased from 25% in March 2002 to 65% in March 2008[1]. The running cost of a digital camera is relatively lower than that of a film camera, and operations of a digital camera is easier than those of a film camera. Today, various types of people take photographs in everyday life, and enormous new photographs arise day by day. So a user has to manage a large amount of photographs. It is one of the most important issues how to manage a lot of personal photographs easily and effectively. On the other hand digital photographs are used for variety kinds of purposes. Effective utilization of a lot of personal photographs is another issue.

Up to now, many techniques for efficient management and effective utilization of personal digital photographs have been reported. Most of the conventional techniques for personal photograph management could be categorized into three types: search, browse and recommendation. The objective of search and browse is

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 623–630, 2009.

to find a photograph that satisfies a requirement of user. When a user searches photographs, an information request of the user is represented explicitly as a query. When a user browses photographs, the user can access photographs navigationally with simple operations. On the other hand, in recommendation a user does not have to give any query and operation. The system proposes some photographs to the user based on access logs and profile of the user. The above three approaches for providing photographs to a user suppose that the user has had one or more information requirements previously. The objective of those approaches is to find one or more photographs, which satisfies an information requirement of the user. We named them as *requirement-based approach.*

Many software and web services based on the requirement-based approaches have been designed based on the assumption that they are used on a personal computer and their objective is to find and provide one or more photographs which could satisfy an information requirement of a user. Figure 1 illustrates an overview of the requirement-based approach.

Today, many kinds of application software and web services for managing personal photographs based on the requirement-based approach have been provided. For example, Flickr, which is one of representative web services for personal photograph management, displays thumbnails of photographs on a map based on their location information and a user can browse them intuitively with easy and simple operations. It is possible to fulfill the requirement such as "I want to see the photograph taken in a location", "I want to know the place where a photograph was taken" and so on. Some digital cameras and cellular phones have GPS function and they can assign location information (latitude and longitude) to a photograph automatically.

However, recent few years, novel types of device such as digital photoframe for managing and utilizing personal photographs have attracted many attentions. The main objectives of a digital photoframe are to store a lot of digital photographs with easy operations and to display them continually. Most of digital photoframe displays stored photographs as a form of slideshow. A slideshow is expected to give a user comfortable feeling or to attract interest of a user. In order to fulfill such expectation, the conventional requirement-based approach is insufficient. This is because the requirement-based approach depends on information requirements of a user. However, a user may have no information requirement when the user watches a slideshow on a digital photoframe.

In this paper, we propose a novel approach for managing and utilizing photographs, which are named *interest-based approach.* This approach is used for composing a slideshow of a target set of personal photographs of a user. Up to now, some techniques for composing a slideshow have been reported. However, in the conventional techniques, photographs are ordered in terms of their file names, shooting date and time or at random. These conventional techniques are inefficient for attracting interest of a user. In our approach, the system supposes that a user has had no information requirement and composes a slideshow of personal photographs that arises interests in the user. We named such a slideshow as *insightful slideshow.* Figure 2 shows an overview of interest-based approach.

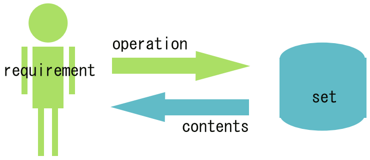


Fig. 1. Overview of Requirement-based approach

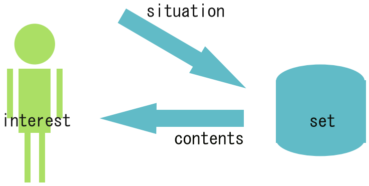


Fig. 2. Overview of Interest-based approach

We suppose that each personal photograph in a target set has a tag manually or automatically. In order to compose an insightful slideshow, the weight of a semantic relation between two photographs is calculated using tags assigned to the photographs and the Web. We suppose that co-occurrence of the tags in the Web reflects relation between the photographs. The photographs are organized into a network and an optimized route in the network is found. The route represents an effective order of photographs as an insightful slideshow.

The composition of this paper is as follows. Section 2 describes some related works. Section 3 show on overview of our approach. Section 4 describes our algorithm for compose an insightful slideshow of stored personal photographs. The summary and future plans are described in Section 5.

2 Related Works

Ba-log[2] manages photographs with the location information where a photograph has been taken. Ba-log provides an interface in which a thumbnail of each photograph is arranged on a map based on its location information. With this interface a user can browse photographs according to their geographical location.

Iwazaki et al.[3] have been proposed a technique for indexing photographs. The geographical location and shooting direction of a photographer when the photographer took a photograph are used for indexing the photograph. Pairs of keyword and location information corresponding to the keyword are stored in a database. The system finds new keywords in the database to a photograph based on its geographic location and shooting direction and recommends some of them relative to the photograph. A user selects suitable keywords from the proposed keywords, and assigned to the photograph as tags. A user can search photographs using assigned tags from various viewpoints.

Fujita et al.[4] has proposed a search interface for photographs based on gaze points of a photographer. In the interface, not only the shooting point but also the gaze point, which is the photographic subject, is considered for browsing photographs. The shooting vector of a photograph is defined as a vector from its shooting point to its gaze point. When a user clicks an object on a map, the clicked point is assumed to be a gaze point. Shooting vectors whose photographs contain the gaze point are searched from the database. The shooting vectors are represented as search results. By selecting a shooting vector in the result, the user can browse photographs about the clicked object.

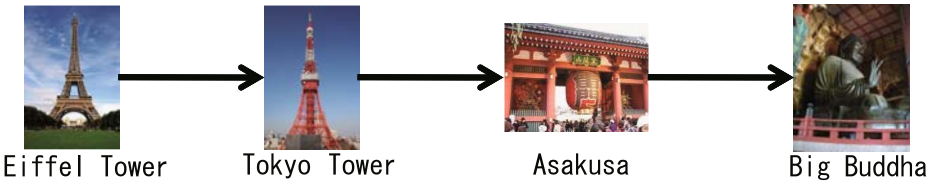


Fig. 3. An Example of Insightful Slideshow

SpaceTag [5] is a system for public information service on the geographic space. A user can make and browse virtual objects on the geographic space with a mobile device. Virtual objects in this system can be accessed in a specific location, date and time.

These systems provide easy and intuitive access of photographs using maps or geographic space. However, they are based on the requirement-based approach, so could not be applicable to attract user's interest.

Jin et al.[6] and Ishida et al.[7] use the Web as knowledge to extract an interpersonal relation network. A relation between persons in a network is estimated based on the co-occurrence of their names on the Web. When the co-occurrence of the names of two persons is high, it is considered that the two persons are closely related. The co-occurrence is derived from the number of web pages in a result of a web search. The weight of a relation is calculated with the Simpson Coefficient. In this paper, we propose a method that derives relations between photographs using the web. This is similar to the method for deriving interpersonal relation network. However, our objective is to compose a slideshow of photographs based on a derived relation network.

3 Insightful Slideshow of Photographs

The objective of our research is to develop a novel method for composing an insightful slideshow of personal digital photographs. An insightful slideshow could attract user's interests. In an insightful slideshow generated by our system, photographs are arranged in a meaningful order. Figure 3 shows an example of insightful slideshow. In this example, a photograph of "Tokyo Tower" is presented continuously after a photograph of "Eiffel Tower" because they contain the same type of buildings. Similarly, a photograph of "Asakusa" is displayed after the photograph of "Tokyo Tower" because they are popular sightseeing spots in Tokyo, and a photograph of "Big Buddha" is displayed after the photograph of "Asakusa" because they are strong related to Buddhism in Japan.

Various type of insightful slideshow of the same set of photographs can be considered. This is because various type of relation can be considered between the same pair of photographs. When new photographs are added into a target set of photographs, another type of slideshow might be organized from a different viewpoint.

4 Composition of Insightful Slideshow

This section describes a technique for deriving semantic relations between photographs, and how to compose an insightful slideshow of the photographs based on their semantic relations.

4.1 Deriving Semantic Relations between Photographs

We suppose that every photograph has been assigned one or more tags manually or automatically. The weight of a relation is calculated based on co-occurrence of tags on the web. We think that the weight of a relation between photographs A and B is high, when the co-occurrence of the tags assigned to A and B is high. Suppose that $T = (t_1, t_2..., t_n)$ is the set of all tags to be assigned to a set of target photographs, and a tag which is assigned to a photograph p is represented as $tag(p)$. The weight of a relation between two photographs p_1, p_2 is calculated using the Simpsons coefficient as formula (1)

$$S(t_1, t_2) = \frac{page(t_1) \cap page(t_2)}{min(|page(t_1)|, |page(t_2)|)} \tag{1}$$

Here, t_i is a tag to be assigned to a photograph, and $page(t_i)$ is the set of all Web pages which contain t_i in their texts. We think that the relation of tags t_i and t_j is strong when the value of $S(t_i, t_j)$ is large. And the relation of the photographs with the tags t_i and t_j is also high when the relation of tags t_i and t_j is strong.

4.2 Construction of Slideshow

Here, we consider a method for composing a slideshow that includes photographs in a target set. Each photograph appears only one time in a slideshow. We assume that continuously presenting two photographs whose weight of relation is high could give a user comfortable impression. We think that a route in a network contains all nodes and edges whose weight of relation is high represents a comfortable presentation order of photographs, where the weight of each edge represents the weight of a relation. We can obtain the most appropriate presenting order of photographs in a slideshow, when we discover the route where the summation of weights of all edges in the route is maxim. However, the problem of finding such route in a network from all the combinations is NP-hard. For answering this problem, we use the nearest neighbor method[8] as an approximate solution method. The following shows our algorithm for deciding a route in a network for an insightful slideshow.

1. We create a network in which each node represents a photograph and each edge represents a relation between two photographs. The edge between node p_1 and p_2 is represented as (p_1, p_2).
2. Calculate of the weight of every relation in the network. The weight of a edge (p_1, p_2) is calculated by means of the Simpsons coefficient $S(tag(p_1), tag(p_2))$.

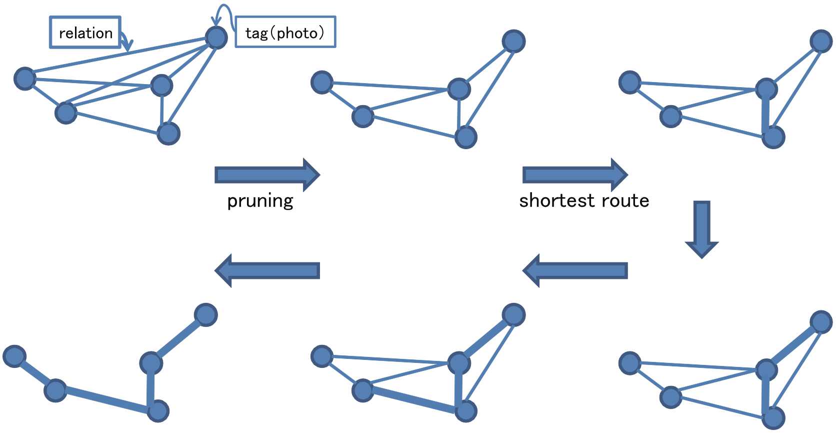


Fig. 4. Overview of Shortest Route Detection

3. For each edge (p_1, p_2), if its weight $(S(tag(p_1), tag(p_2))$ is less than a threshold, it is pruned.
4. All nodes in the network are inserted into a unvisited node list V.
5. For each edge (p_1, p_2), its distance $D(p_1, p_2)$ is calculated by formula (2).

$$D(p_1, p_2) = \frac{1}{S(tag(p_1), tag(p_2))} \tag{2}$$

6. The edge whose distance is minimum is found then the nodes of the edge is set as terminal nodes. And, the nodes are deleted from V.
7. In all the edges from the terminal nodes to a node in V, the edge with the minimum distance is found. And, the nodes of the edge are deleted from V.
8. If V is empty, this process is stopped, otherwise go to step 7.

Figure 4 shows an overview of the process. A circle represents a photograph, a thin line represents a relation, and a thick line represents a path in a selected route.

4.3 Example

This section shows an example of slideshow composition based on our proposed method. Figure 5 shows four photographs to be composed into a slideshow. One tag has been assigned to each photograph. The tags are “Eiffel Tower”, “Tokyo Tower”, “Asakusa” and “Big Buddha”. Table 1 shows the Simpsons coefficient values and the distances for every combination of photographs. The number assigned to an arrow represents the distance between photographs. Figure 6 shows a slideshow of four photographs that is composed by our proposed method. In this figure, an arrow represents an order of displaying two photographs. The number over an arrow represents the distance between two photographs. Intuitive semantic relations can be assumed between each pair of neighboring photographs

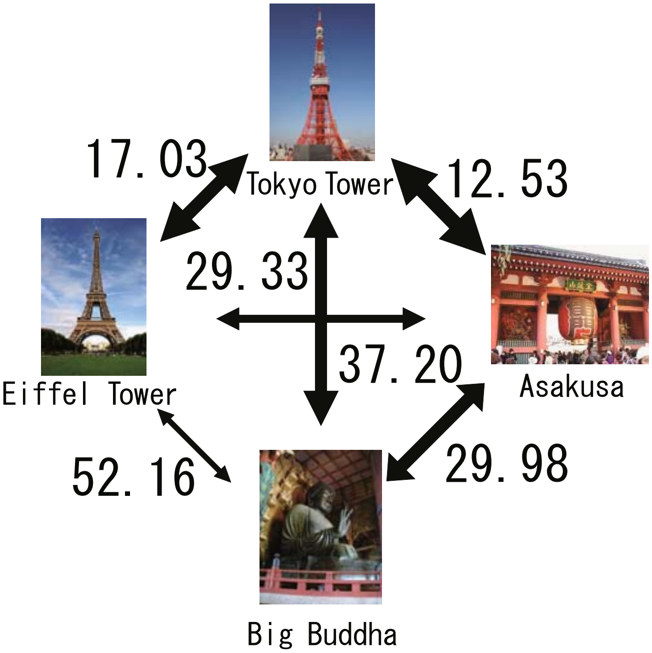


Fig. 5. Example of Photographs and Distances between Them

Table 1. Example of the Simpsons Coefficient Values and Distances between Photographs

rank	tag1	tag2	simpsons	distance
1	Tokyo Tower	Asakusa	0.0798	12.53
2	Eiffel Tower	Tokyo Tower	0.0587	17.03
3	Tokyo Tower	Big Buddha	0.0341	29.33
4	Asakusa	Big Buddha	0.0334	29.98
5	Eiffel Tower	Asakusa	0.0269	37.20
6	Eiffel Tower	Big Buddha	0.0192	52.16

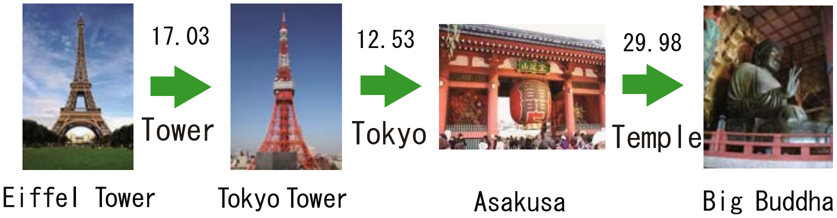


Fig. 6. A Slideshow as a Result of Our Method

in the slideshow. The term under an arrow shows an example of such semantic relation. Therefore it could be expected that the slideshow would give a user a kind of comfortable impression.

5 Conclusion

In this paper, we proposed a method for composing an insightful slideshow automatically. This method provides a slideshow with natural flow by reveal an insight of user's intention.

We have a plan to develop a technique of automatic assignment of tags to a content object by using map information and Web information. Additionally, it is scheduled to evaluate our proposed method by using a large-scale set of photographs.

In this paper, we focus on personal photographs as target content, however the proposed method can be applicable to the other types of digital content. For instance, when a user is watching a news clip, user's personal photographs that have been taken in the place concerned to the news clip. A user could enjoy a new type of information content in which public content and personal content are mixed.

References

1. Cabinet Office.: Survey of consumer behavior. adoption rate (2008), http://www.esri.cao.go.jp/jp/stat/shouhi/shouhi.htm
2. Hiroki, U., Kosuke, N., Tetsurou, T., Ohmukai, I., Hideaki, T.: Ba-log. The Japanese Society for Artificial Intelligence (2004)
3. Kiyoko, I., Kazumasa, Y., Naokazu, Y.: Indexing Photos Based on Shooting Position and Orientation. MIRU (2005)
4. Hideyuki, F., Masatoshi, A., Koji, O.: Retrieval System for Digital Photographs using Observing point of Photographs. DEWS (2003)
5. Ken, M., Megumi, N., Hiroyuki, T., Kambayashi, Y.: Design and Implementation of the SpaceTag Prototype System: An Object System with Spatio-temporarily Limited Access(Special Issue on Knowledge and Information Sharing). Transactions of Information Processing Society of Japan 41(10), 2689–2697 (2000)
6. YingZi, J., Yutaka, M., Mitsuru, I.: Extracting Artist Network from World Wide Web. The Japanese Society for Artificial Intelligence (2006)
7. Keisuke, I., Yutaka, M., Yuki, Y.: Mr. Network Extraction: A Network Extraction System from the Web. The Japanese Society for Artificial Intelligence (2007)
8. Murano, T., Matsumoto, N.: Comparison of Approximate Methods for Traveling Salesman Problem. IEICE technical report. Nonlinear problems 102(625), pp.1–6 (2003)

Web-Based System for Supporting Participation in International Conferences

Akira Hattori, Shigenori Ioroi, and Haruo Hayami

Faculty of Information Technology, Kanagawa Institute of Technology,
1030, Shimo-ogino, Atsugi, Kanagawa, 243-0292, Japan
{ahattori,ioroi,hayami}@ic.kanagawa-it.ac.jp

Abstract. To participate in an international conference, we must complete a series of tasks in accordance with the conference schedule. However, graduate students and researchers who have little or no experience of international conferences often have many difficulties in doing this. We propose a system to support their participation in international conferences. Our proposed system combines three functionalities, knowledge management, workflow management, and schedule management, and it takes advantages of several Web services. Our system associates knowledge, which is composed from the results and deliverables of performed tasks and related know-how, email messages sent by users, and their Web-search histories, with the tasks of conference workflows. In addition, when a workflow is created, the system adds important dates of the conference to the user's Web calendar. We built a prototype system and confirmed that it works properly.

Keywords: International conference, Participation support, Web, Knowledge sharing, Workflow.

1 Introduction

To participate in an international conference for presenting research and examining the latest research findings, participants must complete a series of tasks in accordance with the conference schedule. For example, they have to prepare and submit their paper by the deadline, make airline and hotel reservations, etc. They also receive many comments on their paper from teachers or co-authors and revise it accordingly. This means several people are involved in these tasks.

However, graduate students and researchers who have little or no experience of participating in international conferences often do not know how to fill out a paper submission form and what and when to perform, etc. In addition, if it is their first trip overseas for an international conference, they have the added responsibility of trip preparation as well as paper or presentation preparation. In such cases, they collect information by reading books, searching the Web, or asking someone questions. Through experience, we acquire a large amount of knowledge, such as documents including deliverables and know-how, about participating in an international conference. However, today, most of this knowledge is an individual's

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 631–638, 2009.

personal knowledge. Sharing it among the members of a research group or laboratory can help researchers and graduate students, especially who have little or no experience of international conferences, participate in these conferences.

We propose a knowledge sharing system to support graduate students and researchers who have little or no experience of international conferences by combining three functionalities, knowledge management, workflow management, and schedule management.

2 Related Work

Recently, a number of systems to help researchers participate in international conferences have been put on the Web [1][2]. They collect conference information from mailing lists or the Web and track updates; therefore, they support researchers in finding and catching up on conference information. There are also many books and Websites explaining a series of tasks and the typical schedule of conferences [3][4]. However, these systems and books do not provide an environment for continuously collecting knowledge about conference participation.

A great number of systems, such as Wikis (Wiki Wiki Web) and others using different information sharing platforms, have been developed [5][6]. They are useful for collaboratively building knowledge. However, to participate in an international conference, participants have to complete a series of unique tasks in accordance with the conference schedule. That is why existing information sharing systems are inadequate for supporting participation in an international conference.

Several studies have been made on work process-oriented knowledge management to support the reuse and sharing of processing know-how [6][7][8]. Their systems link the know-how to workflow activities or tasks. As mentioned above, participants usually ask the association or coordinator about the conference, paper submissions, etc. via email. They also search the Web to acquire necessary information. It might be helpful to share sent email messages and Web-search process or used keywords on the tasks necessary for conference participation.

The number of people who are using Web calendar services has been increasing. For example, Google Calendar and Yahoo! Calendar. They mainly use them to manage their to-do lists or schedule. Therefore, we designed our system by taking such services into account.

3 Our Proposed System

3.1 Analysis of Workflow to Participate in an International Conference

In this section, we discuss the process of participating in an international conference. There are two types of tasks that participants have to complete to participate in an international conference. One is related to presentations and the other is related to attendance. The latter type of tasks has to be performed whether or not you make a presentation. These two types of tasks are performed through much the same process [4].

We analyzed our experience based on the points mentioned above. As a result, we recognized the following characteristics of workflow for participating in an international conference:

- The tasks related to attendance include conference registration, arrangement and preparation of the trip, and office procedure at one's organization.
- Different tasks are needed according to researchers' affiliations, positions, etc.
- There are various roles, such as first author and second author.
- Multiple workflow can be in process at the same time.

3.2 Outline of Our System

Figure 1 shows the outline of our proposed system.

The system consists of a database, which stores knowledge, workflow information, conference information and user information, and three management functions to associate the knowledge and information with each other. In addition, it uses Web services such as an email server, Web calendar, and Web search engine. The outline of the functionalities is as follows:

Knowledge Management Function. This function associates knowledge and workflow information with each other and provides knowledge according to users' affiliations or their research groups. It also manages email messages sent by users and their Web-search history as knowledge.

Workflow Management Function. This function creates a workflow (a workflow instance to be more precise), by combining common and additional processes. These two types of processes will be explained in a later section. It changes

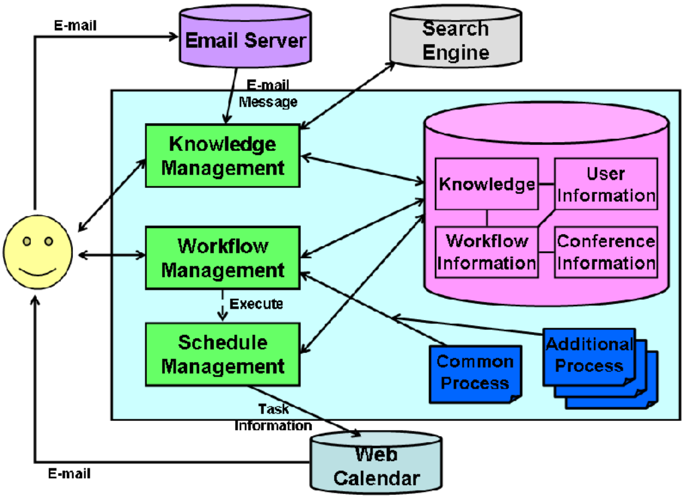


Fig. 1. Outline of our system

the status of the workflow when users input the results of performed tasks, for example a submitted paper or flight schedule, into our system.

Schedule Management Function. This function adds events and tasks to a user's daily Web calendar when he/she starts a workflow. To do this, it uses the conference information stored in the database.

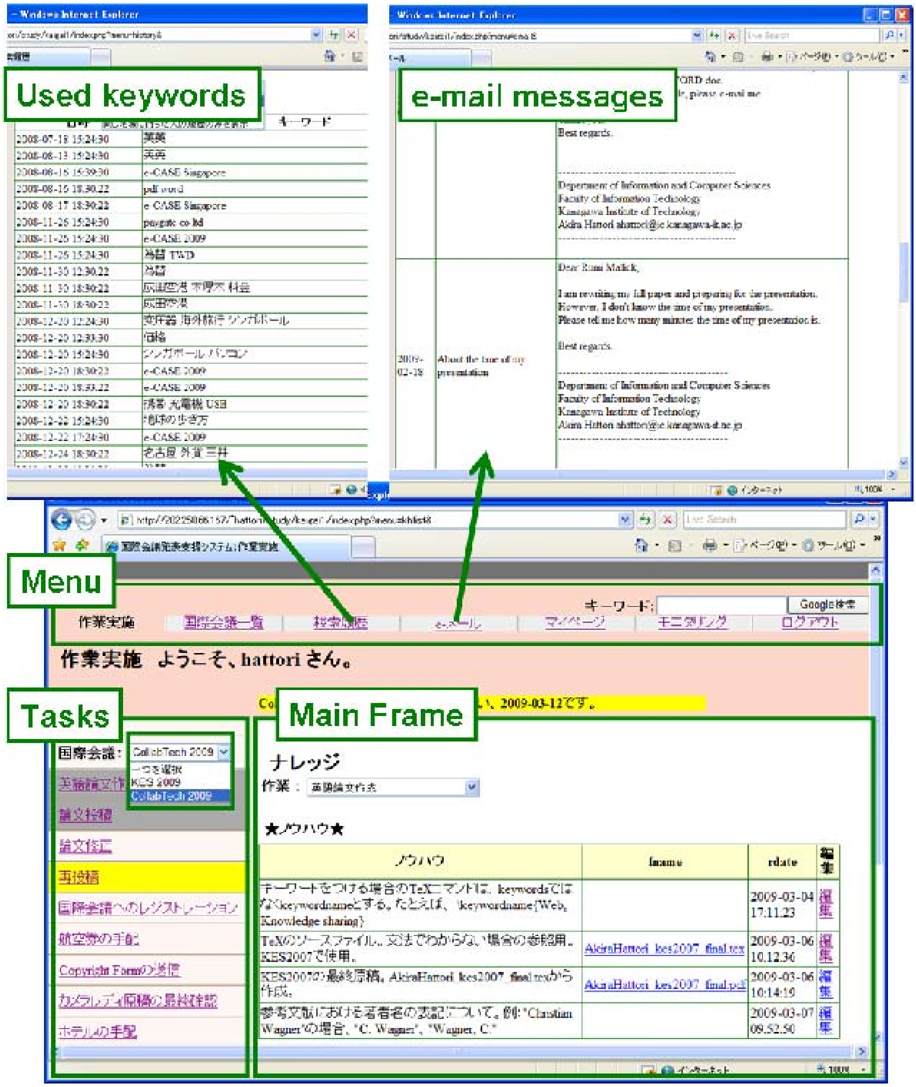
3.3 How to Manage Knowledge

Our system manages the results and deliverables of performed tasks and related know-how, email messages sent by users, and their Web-search histories. The results and deliverables contain documents such as submitted papers, flight schedules, and presentation slides and they are created by completing each of the tasks. In our system, they are inputted on the form corresponding to each task. For example, the system provides the input form for a submitted paper. The know-how is stored by attaching it to a task. It might contain document files and URLs that are useful for successfully completing the task.

As mentioned in the previous section, email messages asking the association or coordinator of a conference, hotel, etc. are worth sharing among users. In our system, when a user sends an email to the system as well as a conference or hotel in his/her usual manner, the system receives the message from the email server, which is Web based, and makes the message available for all users. The message is associated to the workflow of the conference of which the user is going to participate in accordance with the day when it was sent. In addition, because Web-search processes or used keywords for acquiring information related to performed tasks are useful as well, our system manages used search keywords as Web-search histories. Like email messages, the keywords are associated with the workflow using the day when they were used.

3.4 Management of Workflow and Schedule

While a series of tasks for participating in an international conference is performed through much the same process, different tasks are needed according to researchers' affiliations, positions, etc. In our system, therefore, two types of processes are defined. One is common process and the other is additional process. While the former is used regardless of the users' affiliations or positions, the latter depends on them. The system creates a workflow by combining these two processes based on user information. In doing that, conference information is also used to associate the workflow with the corresponding international conference. The system provides the status of the workflow visually based on the conference schedule. The status is changed when users input the results of performed tasks in the manner explained above. In addition, when a user creates a new workflow, our system adds important dates of the conference, such as deadlines for submission and early registration, on his/her Web calendar, which we assume is used daily. By doing this, the system sets notifications by email using the reminder function of the calendar. This is because



日時	キーワード
2008-07-18 15:24:30	英英
2008-08-13 15:24:30	英英
2008-08-15 15:39:30	e-CASE Singapore
2008-08-15 18:30:22	pdf word
2008-08-17 18:30:22	e-CASE Singapore
2008-11-25 15:24:30	paygate co ltd
2008-11-25 15:24:30	e-CASE 2009
2008-11-25 15:24:30	為替 TWD
2008-11-30 12:30:22	為替
2008-11-30 18:30:22	成田空港 京成 料金
2008-11-30 18:30:22	成田空港
2008-12-20 12:24:30	変圧器 海外旅行 シンガポール
2008-12-20 12:33:30	価格
2008-12-20 15:24:30	シンガポール パソコン
2008-12-20 18:30:22	e-CASE 2009
2008-12-20 18:33:22	e-CASE 2009
2008-12-20 18:30:22	携帯 充電機 USB
2008-12-22 15:24:30	地球の歩き方
2008-12-22 17:24:30	e-CASE 2009
2008-12-24 18:30:22	名古屋 外貨 三井

ORD doc.
e, please e-mail me

Best regards.

Department of Information and Computer Sciences
Faculty of Information Technology
Kanagawa Institute of Technology
Akira Hattori ahattori@ic.kanagawa-it.ac.jp

2009-02-18 About the time of my presentation

Dear Rumi Mallick,

I am rewriting my full paper and preparing for the presentation.
However, I don't know the time of my presentation.
Please tell me how many minutes the time of my presentation is.

Best regards.

Department of Information and Computer Sciences
Faculty of Information Technology
Kanagawa Institute of Technology
Akira Hattori ahattori@ic.kanagawa-it.ac.jp

作業実施　ようこそ、hattori さん。

Co… い、2009-03-12です。

ナレッジ

作業：英語論文作成

★ノウハウ★

ノウハウ	fname	rdate	編集
キーワードをつける場合のTeXコマンドは、keywordsではなく\keywordnameとする。たとえば、\keywordname{Web, Knowledge sharing}		2009-03-04 17:11:23	編集
TeXのソースファイル。文法でわからない場合の参照用。KES2007で使用。	AkiraHattori_kes2007_final.tex	2009-03-06 10:12:36	編集
KES2007の最終原稿。AkiraHattori_kes2007_final.texから作成。	AkiraHattori_kes2007_final.pdf	2009-03-06 10:14:19	編集
参考文献における著者名の表記について。例："Christian Wagner"の場合、"C. Wagner", "Wagner, C."		2009-03-07 09:52:50	編集

Fig. 2. User interface of our prototype

graduate students and researchers who have little or no experience of international conferences often do not know what and when to perform.

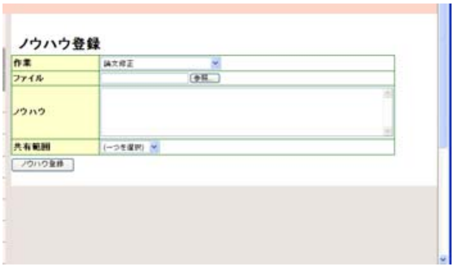
3.5 Prototype System

We developed a prototype of our proposed system. We used Yahoo!Mail, Google Web search, and Google Calendar as the email server, Web search engine, and Web calendar, respectively.

Figure 2 depicts the user interface of the prototype. It largely consists of a menu, a list of tasks, and a main frame. The tasks in the list correspond to the user-selected conference, which is the workflow. They are presented in accordance with the status of the workflow and the conference schedule. When users click a task on the list, the corresponding input form, knowledge related to the task, and a link to a form to attach know-how to the task. The link opens the form like Fig. 3.

Users receive email messages and Web-search histories, which are used keywords, by clicking the item on the menu. This enables our system to access the email server. The menu also contains a Web-search box, in which they can search the Web by inputting keywords. Then the keywords are added to the database as knowledge.

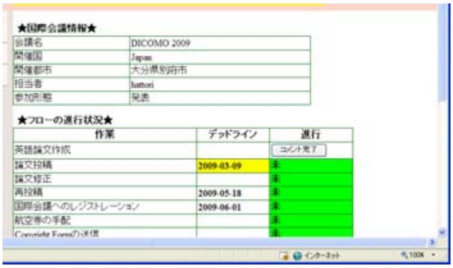
Besides the list of tasks, users can also see the status of the workflow along with the conference schedule in the main frame by clicking that item on the menu. If a user is a second author of a paper, he/she will be given a button to input what he/she has commented about the paper (see Fig. 4).



ノウハウ登録

作業	論文修正
ファイル	
ノウハウ	
共有範囲	(一つを選択)

Fig. 3. Form to attach know-how



★国際会議情報★

会議名	DICOMO 2009
開催国	Japan
開催都市	大分県別府市
担当者	hattori
参加形態	発表

★フローの進行状況★

作業	デッドライン	進行
英語論文作成		コメント完了
論文投稿	2009-03-09	未
論文修正		未
再投稿	2009-05-18	未
国際会議へのレジストレーション	2009-06-01	未
航空券の手配		未
Copyright Formの送信		未

Fig. 4. Status of workflow

4 Discussion

In this section, we discuss the advantages of our proposed system and some improvements that should be made.

First of all, we will take up common situations in the process of paticipating in an international conference. In preparing a paper for an international conference, we often see one that we submitted to the previous conference. The purpose is, for example, to understand how to format the paper. Also, we frequently see a prepared document to complete the same task before. For example, the time when we describe the changes that we have made to satisfy the requirements of the reviewers, etc. Existing process-oriented knowledge management systems only link know-how and documents to workflow activities. Unlike them, our system not only links them to workflow activities, but also organize them by international conferences, that is workflow instances. Therefore, our system is more suitable for the situations than the existing systems.

In addition, when the official language of an international conference is different from the users' native languages, it is important to be able to share input examples of various forms such as registration, inquiry email messages, and keywords used to find information related to performed tasks on a Web search engine, especially for researchers who have little or no experience in conferences. The existing systems manage know-how and documents, but they do not do email messages and Web-search histories. Thus, they are not inadequate for supporting participation in international conferences. On the contrast, our system makes it possible to share all of them.

Moreover, our system adds important dates of an international conference on a user's Web calendar when the workflow of the conferenct is created. As a consequence, the user can always see the dates. In addition, we made active use of the functionality of the calendar to implement our system. This made it possible for us to easily incorporate the reminder function into the system. Without this functionality, the system has to properly manage the settings for notifications and send emails to users at the right time, which is not easy to do. Therefore, there are benefits of using existing calendar services in our system.

We believe records for successfully performing the tasks, for example, the date when a user started a task, are also useful as task-related knowledge. However, our system does not provide the function for sharing such knowledge; therefore, we need to incorporate such a recording function into the system. In addition, it is also necessary to improve the management for sent email messages and used Web-search keywords. Our prototype associates them with the workflows of the conferences in which users are going to participate, but it stores them only in the order in which they were used. If they are associated with tasks based on the conference schedule and the date when they were used, they should be more useful. In that case, used keywords will help users with Web search as the navigation.

By the way, knowledge in our system is composed from the results and deliverables of performed tasks and related know-how, email messages sent by users, and their Web-search histories. However, they are not mutually associated. To make our system more useful, we need to add mechanisms to show the associations.

5 Conclusions

We proposed a system to support graduate students and researchers who have little or no experience of international conferences. Our system combines three functions, knowledge management, workflow, and scheduling. We built a prototype by making effective use of several Web services, and confirmed it works properly.

Future work includes the improvement of the function to manage knowledge. In addition, it is necessary to appropriately deal with different processes among conferences or schedule changes, such as deadline extension. Moreover, we need to carry out quantitative evaluation of our system based on a broad implementation test.

References

1. Brennhaug, K.E.: EventSeer: Testing Different Approaches to Topical Crawling for Call for Paper Announcements (2005)
2. Takada, T., Kanai, H., Nishimura, T.: ConfShare: A Unified Conference Calendar for Researchers. In: ACM Computer Supported Cooperative Work 2008 (2008)
3. Deguchi, M.: Internationa Conference for the First Time. Nippon-Hyoron-sha (2008) (in Japanese)
4. Okada, M.: How to Write a Scientific Paper (in Japanese), http://www.okada-lab.org/Ronbun/
5. Wagner, C., Narasimha, B.: Supporting Knowledge Management in Organizations with Conversational Technologies: Discussion Forums, Weblogs, and Wikis. Journal of Database Management 16(2), 1–8 (2005)
6. Fuchs-Kittowski, F., Kohler, A.: Wiki Communities in the Context of Work Process. In: Proceedings of the 2005 International symposium on Wikis, pp. 33–39 (2005)
7. Holz, H., Rostanin, R., Dengel, A., Suzuki, T., Maeda, K., Kanasaki, K.: Task-based Process Know-how Reuse and Proactive Information Delivery in TaskNavigator. In: Proceedings of the 15th ACM Conference on Information and Knowledge Management, pp. 522–531 (2006)
8. Dustdar, S.: Reconciling Knowledge Management and Workflow Management Systems: The Activity-based Knowledge Management Approach. Journal of Universal Computer Science 11(4), 589–604 (2005)

Analyzing the Relationship between Complexity of Road Networks and Mobile Agents' Simulation

Kazunori Iwata[1], Nobuhiro Ito[2], Yoichi Setoguchi[3], and Naohiro Ishii[2]

[1] Dept. of Business Administration, Aichi University
kazunori@vega.aichi-u.ac.jp
[2] Dept. of Applied Information Science, Aichi Institute of Technology
n-ito@aitech.ac.jp, ishii@aitech.ac.jp
[3] Dept. of Compute Science and Engineering, Nagoya Institute of Technology
agent-staff@phaser.elcom.nitech.ac.jp

Abstract. This paper analyzed the relationship between the evaluation of multi-agent systems and the agents' environments. We define a movement difficulty for maps, using complexity indexes[1] and a vehicle movement simulator. In addition, we report on experiments carried out to confirm whether the movement difficulty can be used to estimate the results of the evaluation of agents. Finally, we investigate the similarity between the results of the simulations and the analysis.

1 Introduction

In a multi-agent system, agents are influenced by their surroundings, which we refer to as the agents' environment. In this paper, we analyze the relationship between the evaluation of mulllti-agent systems and the agents' environments. We focus particularly on multi-agent simulations in which maps are used to depict the agents' environments. In research on agents, it is a challenge to find a method with which to evaluate the behavior of the agents or multi-agent systems, as the evaluation depends on the environments in which the agents exist [2,3]. Hence, it is necessary to clarify the relationship of the evaluation of multi-agent systems and the environments of agents in order to find a method to evaluate multi-agent systems whilst taking account of their environments. Moreover, to clarify the relationship, the environments need to be analyzed and quantified. Previously, we attempted to analyze the relationship between the evaluation of agent systems and the environments of the agents [1] by defining 13 complexity indexes for road networks and analyzing the indexes for several areas. The definitions given in our earlier paper, however, have some issues, which are solved in this paper, using the Variance Inflation Factor(VIF). Furthermore, we define movement difficulty for maps using the complexity indexes and a vehicle movement simulator. Experiments have been carried out using both the dial-a-ride system and RoboCupRescue simulations to confirm whether or not movement difficulty can be used to estimate the results of the agents' evaluation. We also investigate the similarity between the simulation results and those obtained through analysis. Finally, we discuss the results of the investigations.

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 639–647, 2009.

2 Evaluation of Multi-agent Systems

Multi-agent systems refer to systems in which a number of agents interact with one another to solve complicated tasks [4]. In this paper, "agents" refer to actors that are able to identify their situation by interacting with their environments and can then solve many kinds of problems autonomously.

In research on agents, it is a challenge to find a method to evaluate the behavior of agents or multi-agent systems, because there is no explicit evaluation method for agents that transcends the development of agents' theory and the implementation and practical application of agents. However, it is difficult to specify evaluations for agents due to the interdependence of agents and their environments. To illustrate this difficulty consider the following.

(i) If agents *A1* and *A2* are working in an environment *E*, it is easy to evaluate which agent is better suited to the environment. However, the result of the evaluation will not be applicable to other environments except *E*.

(ii) If agent *A1* performs better than agent *A2* in environment *E1*, but agent *A2* performs better than agent *A1* in environment *E2*, it is difficult to decide which agent's performance is better overall.

These examples illustrate why it is necessary to clarify the relationship between agents and their environments in the evaluation of agents and multi-agent systems. To clarify the relationship, a detailed analysis of the environment is required. In this paper, we focus on an agent whose movement is based on maps. In other words, we select an autonomous robot agent, since there are many different kinds of multi-agent systems. Moreover, we concentrate on map data as the environment and analyze it to clarify the relevance. Map data includes various features such as roads, railroads, rivers, buildings, geographical features and so on. Road information is especially noted and analyzed, because there is a strong relation between roads and the evaluation of agents whose movement is based on maps.

3 Maps and a Road Network

3.1 Map Data and Road Information

We use 1/25,000 map data for all the areas in Japan. This data includes roads, railroads, rivers, shorelines of lakes, coastlines, administrative districts, ground control points, place names, public institutions and altitude. These maps are released by the Geographical Survey Institute[5] and the data is expressed in G-XML, which is the Japanese Industrial Standards format. We use the information about roads, which consists of nodes for roads and road edges. Nodes for roads depict intersections, blind alleys and junctions between roads, and have the attributes "ID", "longitude", and "latitude". Road edges are roads, each of which connects the interval between two road nodes. Their data attributes are "ID", "type", "width of road", and the "IDs of the two end points expressed as road nodes".

3.2 A Road Network

In this paper, we regard the information about roads as a road network and define a road network using the attributes given by this information. We represent a road network G, as a weighted digraph (V, E). "V" indicates a point set regarded as a node set in a graph and defined in Definition 1. "E" indicates a road set regarded as an edge set in a graph and defined in Definition 2.

Definition 1. ***Point Set***

$$V = \{v | v \textit{ has ``longitude'' and ``latitude''}\}, \tag{1}$$

where v is a node created by a road node.

Definition 2. ***Road Set***

$$E = \{e | e \textit{ has } v_{head},\ v_{tail},\ \textit{"length" and "width"}, v_{head} \in V, v_{tail} \in V\}, \tag{2}$$

where e is a directed edge from v_{head} to v_{tail} created by a road edge, "length" means the Euclidean distance from v_{head} to v_{tail}, and "width" equals "width of road".

Definition 3. ***Road Network***
A road network G is a weighted digraph (V, E) with two weighting functions l and w. These functions are defined as follows.

$$l : E \rightarrow \mathbf{R}, \tag{3}$$

$$w : E \rightarrow \mathbf{R}, \tag{4}$$

where l derives a real-valued attribute "length" from an edge e and w derives a real-valued attribute "width" from an edge e.

A road network G is in the X-Y Cartesian coordinate system in which the x- and y-axes represent latitude and longitude, respectively. Each road e in G is a segment of line connecting a v_i and v_j ($v_i, v_j \in V$ and $v_i \neq v_j$). In addition, G has the following properties.

(i) Strongly-connected digraph.
(ii) No loop back edge.
(iii) No multiple edges between any pair of nodes.

4 Analysis of Road Networks

4.1 Earlier Study

In an earlier study on the analysis of road networks [1], we defined 13 complexity indexes for a road networks and analyzed the indexes for several areas. The complexity indexes are given below.

(1) The average length of all edges.
(2) The average "distance of shortest paths" for all pairs of nodes.

(3) The available roads rate. The available road indicates a wide road with a width is greater than $5.5m$.
(4) The available length rate in roads. This is the ratio of available roads to the total length of all roads.
(5) The available area rate in roads, which indicates the ratio of the area of available roads to the total area of all roads.
(6) The average width of all edges.
(7) The average "maximum flow"[6].
(8) The ratio of intersections to all edges. An intersection means an edge having more than 3 in-degrees and 3 out-degrees.
(9) The ratio of arranged area, which indicates similarity between a graph G and a square grid.
(10) The ratio of straight edge. A straight edge indicates almost a straight path.
(11) The ratio of straight edge to distance.
(12) The ratio of passing through a road. This indicates the total number of times an edge is used by all shortest paths.
(13) The average local road-connectivity. The local road-connectivity of two edges $m, n \in V$ is the smallest edge cut disconnecting m from n.

We performed evaluation experiments, with an agent moving like a vehicle on a map. We also investigated the similarity between the experimental results and those obtained through analysis.

4.2 Issues and Concerns

Based on our earlier studies we are able to specify the maps on which an agent can move effectively. However, the complexity indexes are occasionally less reliable because some of the indexes have a high degree of interdependence among each other. (This is referred to as multicollinearity in multiple regression analysis.) Moreover, the earlier studies considered agents' movement, but not agents' evaluation.

In this paper, we solve the multicollinearity using a Variance Inflation Factor (VIF) and investigate the relationship between the complexity indexes and agents' evaluation in both the Dial-a-Ride System Simulation[7] and RoboCupRescue Simulation[8,9].

4.3 Selection of Complexity Indexes

We solve the multicollinearity of the complexity indexes using VIF.

Let $x_1, x_2, \ldots, x_i, \ldots x_{n-1}, x_n$ be explanatory variables, then the VIF of x_i is calculated as follows.

$$VIF_i = \frac{1}{1 - R_i^2} \tag{5}$$

where R_j^2 is a multiple correlation coefficient between the variable x_i and other variables. If VIF_i is greater than 10, variable x_i has multicollinearity when calculating the multiple regression analysis.

Table 1. Results of the Selection

Complexity Index	VIF
Average "distance of shortest paths" for all pairs of nodes (c_1)	1.8447
Average width of all edges (c_2)	4.5671
Available roads rate (c_3)	7.6851
Available length rate in roads (c_4)	7.5072
Ratio of arranged area (c_5)	4.0125
Ratio of straight edge to distance (c_6)	4.0125
Ratio of passing through a road (c_7)	2.2853
Average local road-connectivity (c_8)	3.4626

The complexity indexes are reduced by iteration through the following steps:

(1) Calculate VIF_i for all existing variables.
(2) Remove x_{max} from the set of variables, if the maximum VIF_{max} is greater than 10. Otherwise, terminate the reduction
(3) Return to step (1).

The outcome of executing these steps is shown in Table 1, where a variable for each complexity index is denoted by $c_i (i = 1, 2, \ldots, 7, 8)$.

5 Relationship between Complexity of Road Networks and Agents' Movement

Having implemented a vehicle movement simulator, we define movement difficulty (MD) on a map using the complexity indexes and the simulator. In order to define MD, we use $c_1 \sim c_8$ as explanatory variables and an average migration time on a map as the dependent variable. The average migration time on a map in the vehicle movement simulator is calculated as follows.

$$t_{avg} = \frac{\sum_{u,v \in V, u \neq v} t_{u,v}}{{}_{|V|}\mathrm{C}_2} \tag{6}$$

where $t_{u,v}$ is the migration time from u to v (or from v to u). t_{avg} indicates a case of mobility on a map; the higher the value of t_{avg} the more difficult it is for a vehicle to move.

The definition of MD is derived from multiple regression analysis with $c_1 \sim c_8$ as explanatory variables and the average migration time on a map as the dependent variable and is given as Definition 4.

Definition 4. ***Movement Difficulty***

$$\begin{aligned} MD(= \widehat{t_{avg}}) = {} & 3.47e^{-5} \times c1 + 0.00919 \times c2 - 0.0405 \times c3 - 0.748 \times c4 \\ & -0.222 \times c5 + 4.01 \times c6 + 0.0399 \times c7 + 0.0307 \times c8 + 0.0697 \end{aligned} \tag{7}$$

where $\widehat{t_{avg}}$ indicates an estimated value for t_{avg}.

Table 2. Effectiveness of MD

	Average Migration Time	Standard Deviation	Mean Absolute Errors
Learning Maps	0.2063	0.07264	0.009673
Test Maps	0.2112	0.08341	0.022500

To test the effectiveness of MD, we use the mean absolute errors for the learning maps and the test maps. The learning maps are used in the derivation of Definition 4 and then the values of MD in the test maps are estimated using the definition. The results of the test are given in Table 2, and show that using this definition, we are able to estimate MD for maps.

6 Evaluation Experiments

6.1 Dial-a-Ride System Simulation

Dial-a-ride is a bus system that operates like a taxi and includes the following processes.

(1) A passenger calls the bus control center and states a destination.
(2) The center re-plans the route of an appropriate bus to service the request.

To evaluate the dial-a-ride system simulation, we focus on usability and profitability. With regards to usability, we specifically address the primary purpose of the bus system, that is to provide a way for passengers to reach their destinations as quickly as possible. From this point of view, usability is defined as follows.

Definition 5. ***Usability***: *Average of the difference between the desired time and actual time when a passenger gets off a bus.*

The profit (or less) of a bus company depends on maintenance, fuel and labor costs, and fare income, which vary with social economic conditions. In addition, fare-pricing causes secondary social effects, which in turn affect the number of passengers. Hence, it is difficult to quantify profitability directly. Instead, we simplify it as the balance between fare revenue and costs, where revenue and costs change in proportion to the number of passengers and buses, respectively. In other words, profitability is defined as follows.

Definition 6. ***Profitability***: *The number of requests occurring in a unit period per bus.*

6.2 Results of Evaluation Experiments for Dial-a-Ride System Simulation

The results of regression analysis with MD as the explanatory variable and usability(U) as dependent variables are shown in Eq. (8) and Fig. 1. The results

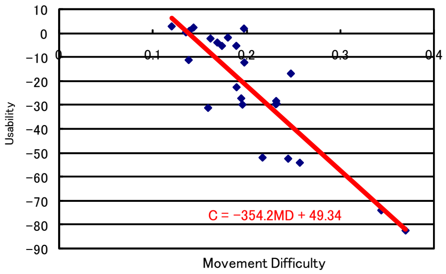


Fig. 1. Results for Usability

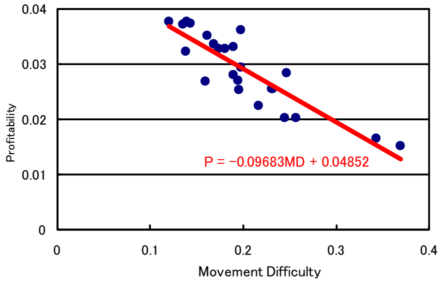


Fig. 2. Results for Profitability

Table 3. Correlation Coefficient and Adjusted R^2 for Dial-a-Ride System Simulation

	Correlation Coefficient	Adjusted R^2
Usability	-0.8608	0.7409
Profitability	-0.8482	0.7195

for MD and probability(P) are shown in Eq. (9) and Fig. 2. The correlation coefficient and adjusted coefficients of determination (written in adjusted R^2) for each evaluation are given in Table 3.

$$U = -106.2MD + 5.805 \tag{8}$$

$$P = -0.04460MD + 0.03957 \tag{9}$$

Table 3 shows that the greater a map's MD is, the lower the usability and profitability are. These results confirm that Eqs. (8) and (9) can estimate the usability and profitability for a map.

6.3 RoboCupRescue Simulation

The RoboCupRescue Simulation is used as a testbed environment and simulates, on a network of computers, a great earthquake and various kinds of disaster-relief activities by multi-agents in a virtual city. The evaluation of the agents in this simulation is given as the “city values retention rate” (written as V_{rate}).

6.4 Results of Evaluation Experiments for RoboCupRescue Simulation

The results of the regression analysis with MD the explanatory variable and V_{rate} as the dependent variable is shown in Eq. (10) and Fig. 3. The correlation coefficient and the adjusted R^2 are shown in Table 4.

$$V_{rate} = -0.31642MD + 0.8999 \tag{10}$$

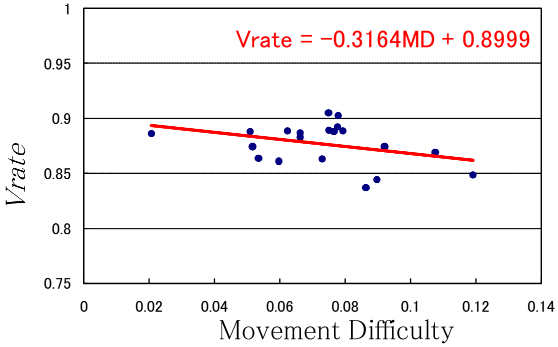


Fig. 3. Results of Regression Analysis

Table 4. Correlation Coefficient and Adjusted R^2 for RoboCupRescue Simulation

	Correlation Coefficient	Adjusted R^2
V_{rate}	-0.3.698	0.1295

Table 3 shows that the greater the map's MD, the slightly lower is the V_{rate}. However, the accuracy of Eq. (10) is not adequate, because V_{rate} is influenced not only by road networks, but also by the number of buildings, the allocation of the buildings and ignition points.

7 Conclusion and Future Work

This paper analyzed the relationship between the evaluation of multi-agent systems and the environments of the agents. We improved the complexity indexes for road networks from our earlier study and defined movement difficulty for maps using the complexity indexes and a vehicle movement simulator. In addition, using regression analysis, we showed the relevance of the movement difficulty in both the dial-a-ride system simulation and RoboCupRescue simulation. We then investigated the results of the simulations and the analysis. These results indicate that movement difficulty is used effectively in simulations where an agent's evaluation is influenced by a road network.

Our future work includes

(1) investigating other environments, besides a road network, for multi-agent systems, and
(2) suggesting ways to evaluate multi-agent systems taking account of their environments.

Acknowledgement

This research was supported by Grant-in-Aid for Scientific Research (No. 19700032, 21700177) awarded from the Japan Society for the Promotion of Science.

References

1. Iwata, K., Ito, N., Kaneda, Y., Ishii, N.: Complexity of road networks as agents' environments. In: Nguyen, N.T., Jo, G.-S., Howlett, R.J., Jain, L.C. (eds.) KES-AMSTA 2008. LNCS, vol. 4953, pp. 474–484. Springer, Heidelberg (2008)
2. Kinoshita, T., Sugawara, K.: Agent Oriented Computing Foundation and Application of Agents (in Japanese). Soft Research Center (1995)
3. Pynadath, D.V., Tambe, M.: Multiagent teamwork: Analyzing the optimality and complexity key theories and models. In: Proceedings of International Joint Conference on Autonomous Agents and Multi-Agent Systems(AAMAS), pp. 873–880 (2002)
4. The Japanese Society for Artificial Intelligence (ed.): Encyclopedia of Artificial Intelligence in Japanese. Kyoritsu Shuppan Co., Ltd. (2005)
5. GSI: 1/25,000 map information of Japan (2003), `http://sdf.gsi.go.jp/` (in Japanese)
6. Wikipedia: Flow network (2007), `http://en.wikipedia.org/wiki/Flow_network`
7. Iwata, K., Ito, N., Noda, I., Ishii, N.: Design and implementation of a simulator for evaluating dial-a-ride systems. In: Hamza, M.H. (ed.) IASTED Conf. on Software Engineering and Applications, pp. 442–448. IASTED/ACTA Press (2004)
8. Robocup rescue home, `http://www.robocuprescue.org/`
9. Tadokoro, S., Kitano, H.: RoboCup-Rescue Technical Committee. In: The RoboCup Federation, RoboCup Japanese National Comittee (eds.): RoboCup Rescue. Kyoritsu Shuppan Co. Ltd. (2000) (in Japanese)

Multi-base Station Placement for Wireless Reprogramming in Sensor Networks

Aoi Hashizume, Hiroshi Mineno, and Tadanori Mizuno

Shizuoka University, 3–5–1 Johoku, Naka–ku, Hamamatsu, Shizuoka 432-8011, Japan
aoi@mizulab.net, {mineno,mizuno}@inf.shizuoka.ac.jp

Abstract. Reprogramming sensor nodes is an effective way of improving wireless sensor networks. The latest reprogramming protocols use radio communication to distribute software data. Although several base stations are optimally placed to disseminate software rapidly in large-scale sensor networks, the performance of reprogramming protocols for multi-base station environments has not been discussed. This paper discusses our evaluation of the features of software dissemination by multi-base station sensor networks. Simulations revealed that the placement and number of base stations were the key parameters in software dissemination.

1 Introduction

The recent advances in MEMS (Micro Electro Mechanical Systems) and low-power wireless communication technology have led to the development of wireless sensor networks (WSNs). A typical WSN consists of a number of small battery-powered sensor nodes that autonomously sense, collect, and transfer various data. There are many WSN applications and services, including structural monitoring, security, and position tracking. These applications are expected to improve our daily lives and create an intelligent society. Most WSNs include state-of-the-art technologies (e.g., ad-hoc network routing, data processing, and position estimation), and these technologies are still developing. Therefore, their codes will be modified or extended in the future for long-running applications using WSNs. Thus, a method to efficiently reprogram many deployed sensor nodes is necessary.

Recently, wireless reprogramming has been extensively researched [1] – [4]. Wireless reprogramming distributes new code to numerous sensor nodes using wireless multihop communication. A base station disseminates large amounts of program data to the entire network. As WSNs are becoming increasingly larger, more than one base station is needed to disseminate software. However, most reprogramming protocols that have been discussed assume a single-base station environment, and multi-base station environments have not been considered.

Here, we present methods of placing base stations in large-scale WSNs.

This paper is organized as follows. Section 2 explains some issues related to wireless reprogramming. An overview of our proposed technique, packing, is introduced in Section 3. We describe the simulation environment and evaluate

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 648–655, 2009.

packing in Section 4. Finally, Section 5 summarizes the paper and mentions future work.

2 Related Issues

2.1 Challenges of Reprogramming

Many wireless reprogramming protocols share various design challenges. Here, we deal with the three most important ones [1]:

- **Completion time:** The completion time for reprogramming affects services using WSNs. When we reprogram the network, we have to stop services and wait until code updating has been completed. Thus, we have to minimize the completion time for reprogramming.
- **Energy efficiency:** Sensor nodes are usually battery powered and the sensor node battery provides the energy used in reprogramming. This battery also supplies energy for computing, communication, and sensing functions. Therefore reprogramming must be energy efficient.
- **Reliability:** Reprogramming requires the new code to be delivered throughout the entire network, and the delivered code must be correctly executed on the sensor node.

In the next section, we discuss how we dealt with the two techniques used in several reprogramming protocols to resolve these three challenges.

2.2 Approaches to Reprogramming

Pipelining. Pipelining was developed to accelerate reprogramming in multihop networks [5] [6]. A program in pipelining is divided into several segments, and each segment contains a fixed number of packets. Instead of receiving the whole program, a node becomes a source node after receiving only one segment. Figure 1 shows the process of software distribution in pipelining. There are five sensor nodes deployed linearly in the figure. The dashed lines represent the interference range, and the solid arrows represent the reliable communication range. To avoid the hidden terminal problem, parallel data should be transferred with a spacing of at least three hops. In the figure, while D is sending segment 1 to E, A is simultaneously sending segment 2 to B. Thus, pipelining can transfer program codes fast by overlapping the segments.

Negotiation. Negotiation is used to avoid data redundancy and improve reprogramming reliability. As previously explained, pipelining is done through segmentation. After this, it is necessary to avoid broadcast storms that are caused by dealing with too many segments. A negotiation scheme was developed in SPIN [7]. This scheme uses three-way handshakes between senders and receivers. Figure 2 shows a three-way handshake. There are three types of messages (ADV, REQ, and DATA) in simple negotiation protocols. First, the source node (A)

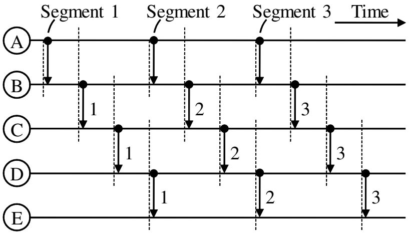


Fig. 1. Three pipelining segments in four-hop network

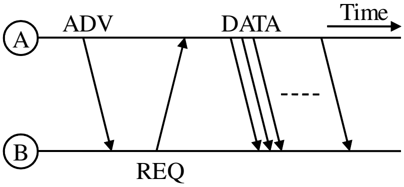


Fig. 2. Three-way handshake

sends an ADV message, which includes its received segment information, to neighboring nodes (B). Second, if the destination node receives the ADV message, it compares its own segment with the received segment information and determines whether it needs the segment sent by the source node. If the segment is needed, the receiver requests the segment from the source node by sending an REQ message. Finally, if the source node receives the REQ message from the destination node, it forwards a requested DATA message. By using this scheme, the source node knows which segment has been requested before sending it out. As a result, data redundancy is reduced.

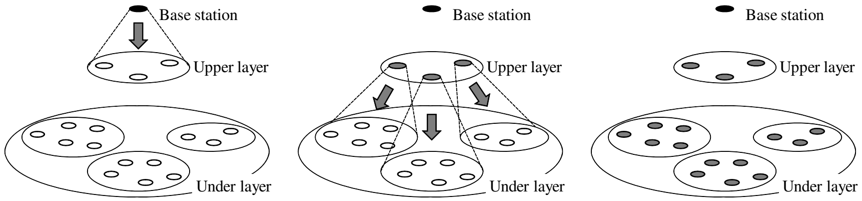
2.3 Hierarchical Reprogramming

Hierarchical reprogramming has been developed to accelerate software distribution and reduce the number of control packets. Firecracker [8] and Sprinkler [9] are known as hierarchical reprogramming protocols. Figure 3 visually depicts their operation. First, the base station sends program codes to nodes in the upper layer of the node hierarchy (i.e., pseudo-base stations). Pseudo-base stations then distribute codes to other nodes in their local areas. Except for Firecracker or Sprinkler, most reprogramming protocols start distributing software from a single base station in the network and assume no hierarchy.

Hierarchical reprogramming protocols improve the efficiency of reprogramming, but no methods of determining where the pseudo-base stations are placed have been discussed. If base stations are deployed in suitable places for reprogramming, the efficiency of reprogramming should be greatly improved.

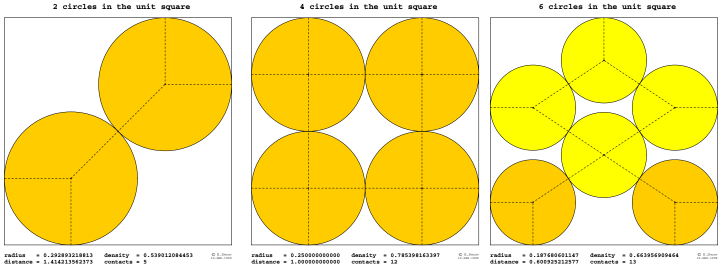
3 Proposed Approaches

Here, we assume that program codes are disseminated at constant speeds from all base stations in a concentric fashion in a planar network. Then, it is the order of placement that minimizes interference between the propagating waves due to software distribution from the base stations. In other words, it is necessary for placement to maximize the dimensions of concentric circles that center on each base station when propagating waves make contact with each other. Then, the method of determining placement is replaced by a circle-packing problem. This



(a) Reprogram upper layer (b) Reprogram under layer (c) Completion

Fig. 3. Example hierarchical reprogramming protocol behavior


2 circles in the unit square
radius = 0.292893218813 density = 0.539012084453
distance = 1.414213562373 contacts = 5
© E. Specht 10-SEP-1999
4 circles in the unit square
radius = 0.250000000000 density = 0.785398163397
distance = 1.000000000000 contacts = 12
© E. Specht 10-SEP-1999
6 circles in the unit square
radius = 0.187680601147 density = 0.663956909464
distance = 0.600925212577 contacts = 13
© E. Specht 10-SEP-1999

(a) 2 circles (b) 4 circles (c) 6 circles

Fig. 4. Packings of equal circles in unit square

problem has been discussed both as a theoretical geometrical problem as well as a hard test for global methods of optimization. The circle-packing problem is replaced by one that maximizes the minimum value of the distance between two circles and the distance between the circle and the boundary of the unit square. By using several symbols, an optimized solution to the problem is expressed as follows [10].

$$max \quad min[\min_{(i,j)\in D_n} d(x_i, x_j), \min_{i=1,\dots,n} d(x_i, SB)]$$
$$s.t. \quad x_i \in S, \quad i = 1, ..., n$$

S: Unit square
SB: Boundary of S
n: Number of equal circles that are packed in S
x_i: Center of number i circle
D_n: Delaunay triangulation with $x_i(i = 1, ..., n)$
$d(x_i,\ x_j)$: Euclidean distance between x_i and x_j
$d(x_i,\ SB)$: Euclidean distance between x_i and SB

The best known packings of equal circles in a unit square are already known. Figure 4 shows some examples of circle packings [11]. Thus, we propose multi-base station placement using a packing approach. In this approach, base stations are placed in the center of each circle.

We developed two other methods called the random and the edge approach to evaluate the packing approach discussed in the next section.

- **Random approach:** Placement of base stations is determined by using uniform random numbers.
- **Edge approach:** Base stations are placed at the edges of networks. Additionally, assuming networks are dealt with using Voronoi diagrams, all the dimensions of base-station Voronoi cells are nearly equal.
- **Packing approach:** Base stations are placed at the centers of circles that are packed in the unit square.

4 Simulation Evaluation

4.1 Environments

This section describes our evaluation of the packing approach using the TinyOS [12] network simulator (TOSSIM [13]). The purpose of this evaluation was to prove that the packing approach is superior to other approaches in terms of completion time, network traffic, and power consumption.

First, we will explain the implementation of the simulation and the items we evaluated in it. We adopted MNP [6] as a reprogramming protocol to evaluate the proposed approaches. This state-of-the-art protocol includes pipelining and negotiation. The completion time and network traffic were observed in TOSSIM, but the battery or power consumption of sensor nodes was not duplicated. We then calculated power consumption on the basis of the typical power consumption of Mica2 Motes in Table 1 [1] [14]. When node i sends S_{it} packets and receives R_{it} packets during t ms, the power consumption of node i $P_i(t)$ is expressed as

$$P_i(t) = 20 \cdot S_{it} + 8 \cdot R_{it} + 1.25 \cdot t. \tag{1}$$

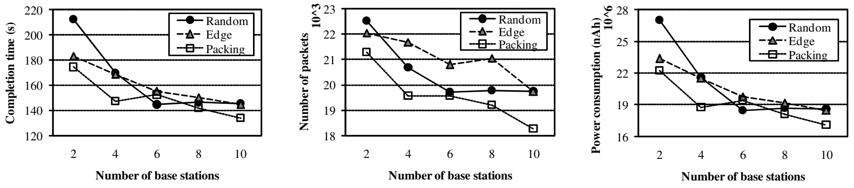
If there are k nodes in the network and the protocol needs T ms to reprogram the whole network, the power consumption of entirety P_{total} is

$$P_{total} = \sum_{i=1}^{k} P_i(T). \tag{2}$$

Next, we will describe the simulation environment. We assumed each node had a transmission radius of 50 feet, so that they could receive messages within a

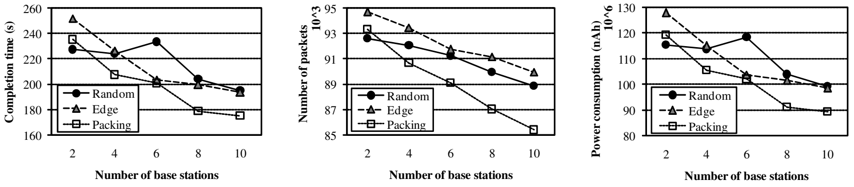
Table 1. Typical power consumption of Mica2 Motes

Operations	Power consumption (nAh)
Send one packet	20.000
Receive one packet	8.000
Idle listen for 1 ms	1.250



(a) Completion time (b) Network traffic (c) Power consumption

Fig. 5. Comparison of three approaches in 10 x 10 network



(a) Completion time (b) Network traffic (c) Power consumption

Fig. 6. Comparison of three approaches in 20 x 20 network

25-feet radius. Nodes were deployed in a reticular pattern of 10 x 10 or 20 x 20. Each node had 40 feet of spacing. Program code in eighty packets was divided into 10 segments and distributed.

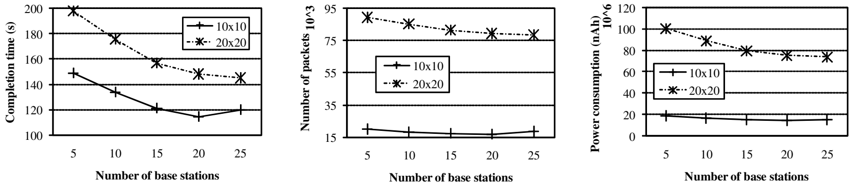
4.2 Base Station Placement

We compared three approaches in two different-sized networks. Figure 5 plots the simulation results in a 10 x 10 network. These graphs confirm all three approaches could be used to shorten the completion time and reduce network traffic and power consumption by increasing the number of base stations. The packing approach had the shortest completion time, least number of packets, and lowest power consumption, irrelevant of the number of base stations.

Figure 6 plots the simulation results in a 20 x 20 network. In the same way as seen in Figure 5, all three approaches can improve reprogramming efficiency. In addition, packing is again the most competent approach to placement.

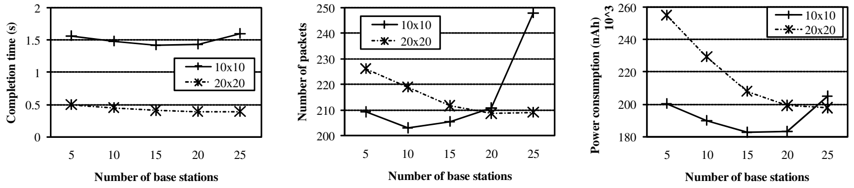
4.3 Network Size and Number of Base Stations

Next, we evaluated the relation between the number of base stations in relation to network size. We only used the packing approach and increased the number of base stations from 5 to 25. Figure 7 compares the performance of an entire 10 x 10 and an entire 20 x 20 network. All results indicate that the characteristics of completion time, number of packets, and power consumption monotonically decrease as the number of base stations increases in a 20 x 20 network. However,



(a) Completion time (b) Network traffic (c) Power consumption

Fig. 7. Comparison of 10 x 10 network with 20 x 20 network using packing approach (entire network)



(a) Completion time (b) Network traffic (c) Power consumption

Fig. 8. Comparison of 10 x 10 network with 20 x 20 network using packing approach (per node)

in a 10 x 10 network, increasing the number of base stations produced poor results.

Figure 8 plots the performance per node in the two networks. The values in Figure 7 were divided by the number of nodes in each network to create the values in Figure 8. We can see that the 10 x 10 network's reprogramming efficiency clearly deteriorates when the number of base stations is increased. Thus, it is conceivable that there are upper limits to the number of base stations in each network that can help to efficiently improve reprogramming. If so, exceeding the upper limit has an adverse effect on reprogramming. This adverse effect has roots in the features of pipelining. Pipelining works well when there is a large number of hops between base stations and farthest destination nodes. In contrast, a small number of hops causes delay in code distribution. In the 10 x 10 network in Figure 7 or 8, when there are 5 base stations, each base station have to send segments to the farthest four-hop nodes. Further each base stations only has to send segments one-hop nodes when there are 25 base stations. Therefore, increasing the number of base stations shortened the required number of hops to reprogram the entire network and caused inefficient pipelining.

5 Summary and Future Work

We presented our packing approach, which can be used to reprogram large-scale sensor networks efficiently by increasing the number of base stations and placing them in optimal order. The simulations revealed that the packing approach could effectively shorten the completion time and reduce network traffic and power consumption. In addition, we found that we had to determine the number of base stations according to the size of the network.

In future work we intend to change the simulation topologies and evaluate these in various environments, for instance, by scaling network size, placing various barricades, or dealing with node irregularities. We then intend to search for the numbers of base stations and placements that are best suited to the targeted networks.

References

1. Wang, Q., et al.: Reprogramming wireless sensor networks: challenges and approaches. IEEE Network 20(3), 48–55 (2006)
2. Stathopoulos, T., et al.: A remote code update mechanism for wireless sensor networks. CENS, Tech. Rep. (2003)
3. De, P., et al.: ReMo: An Energy Efficient Reprogramming Protocol for Mobile Sensor Networks. In: Proc. IEEE PerCom, pp. 60–69 (2008)
4. Huang, L., Setia, S.: CORD: Energy-efficient Reliable Bulk Data Dissemination in Sensor Networks. In: Proc. IEEE INFOCOM, pp. 574–582 (2008)
5. Hui, J.W., Culler, D.: The Dynamic Behavior of a Data Dissemination Protocol for Network Programming at Scale. In: Proc. ACM SenSys, pp. 81–94 (2004)
6. Kulkarni, S.S., Wang, L.: MNP: Multihop Network Reprogramming Service for Sensor Networks. In: Proc. IEEE ICDCS, pp. 7–16 (2005)
7. Kulik, J., et al.: Negotiation-Based Protocols for Disseminating Information in Wireless Sensor Networks. Wireless Networks 8(2/3), 169–185 (2002)
8. Levis, P., Culler, D.: The Firecracker Protocol. In: Proc. ACM SIGOPS European Workshop (2004)
9. Naik, V., et al.: Sprinkler: A Reliable and Energy Efficient Data Dissemination Service for Wireless Embedded Devices. In: Proc. IEEE RTSS, pp. 277–286 (2005)
10. Nanzan University: Academia, Mathematical Sciences and Information Engineering, Vol. 2, pp. 55–60 (2002)
11. Packing of circles in the unit square, http://hydra.nat.uni-magdeburg.de/packing/csq/csq.html
12. TinyOS Community Forum, http://www.tinyos.net/
13. Levis, P., et al.: TOSSIM: Accurate and Scalable Simulation of Entire TinyOS Applications. In: Proc. ACM SenSys (2003)
14. Crossbow Technology, Inc.: Mica2 Wireless Measurement System Datasheet. (2003)

Optimization of Transport Plan for On-Demand Bus System Using Electrical Vehicles

Kousuke Kawamura[1] and Naoto Mukai[2]

[1] Dept. of Electrical Engineering, Graduate School of Engineering,
Tokyo University of Science
Kudankita, Chiyoda-ku, Tokyo, 102-0073, Japan
j4309623@ed.kagu.tus.ac.jp

[2] Dept. of Electrical Engineering, Faculty of Engineering, Division 1,
Tokyo University of Science
Kudankita, Chiyoda-ku, Tokyo, 102-0073, Japan
mukai@ee.kagu.tus.ac.jp

Abstract. An on-demand bus system is now attracting attention as an alternative transport system for traditional fixed-route bus in Japan. In the on-demand bus system, buses transport customers door-to-door according to users' demands, a user can freely specify the position of bus stop in its service area, and the desired time to get the buses. In this paper, we propose a model of the on-demand bus system using electrical vehicles and evaluate its feasibility by computer simulation. The characteristics of the electric vehicles are not considered in the past researches for on-demand bus problem. The improper charge timing decreases the acceptable rate of demands, and the lack of battery charge may occur while the vehicle is moving. In order to avoid such problems, we adopt the genetic algorithm to optimize transport plans. Simulation results showed that our transport model succeeded in the reduction of carbon-dioxide emissions by 80% and the running cost by 60% compared with traditional systems.

1 Introduction

Recently the closing of bus business has been a serious problem in Japan. This problem is caused by the increasing of private automobile possessions and the lack of depots and lines for fixed-route buses. On the other hand, the demand-bus system [1,2] is now attracting attention as a new transport facility. In the on-demand bus system, buses transport customers door-to-door according to users' demand, a user can freely specify the position of bus stop in its service area, and the desired time to get the buses. This study aims to build a new model of transport system using electrical vehicles based on the demand-bus system in an environmentally friendly way.

In order to introduce electrical vehicles to the demand-bus system, it is necessary to construct a transport plan depending on the battery charging for electrical vehicles. There is a past work [3] that solves a route optimization problem

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 656–663, 2009.

for the on-demand bus system by two algorithms, Node Insert Algorithm (NIA) and Genetic Algorithm (GA). However, the work did not consider the transport schedule for electrical vehicles. Therefore, we propose a scheduling method based on NIA and GA for electrical vehicles. In our method, if a user sends a request to the system, a demand for the user is assigned to a vehicle by NIA as an initial transport plan. Then, the GA is applied to the initial transport plan, i.e., crossover and mutation operations optimize the transport plan in consideration of battery power and customer's condition (e.g., time limit). We evaluated our transport model by computer simulation, and the results show that our model succeeded in the reduction of carbon-dioxide emissions by 80% and the running cost by 60% compared with traditional systems.

The remainder of this paper is as follows: Section 2 summaries the model of the on-demand bus system using electrical buses. Section 3 proposes an optimization method for the on-demand bus system using electrical vehicles. Section 4 reports our experimental results. Finally, Section 5 describes conclusions and future works.

2 Model of On-Demand Bus System Using Electrical Vehicles

The demand bus problem is a variant of Vehicle Routing Problem (VRP) [4,5]. This section shows the model of the on-demand bus system using electrical vehicles on the basis of the VRP.

A demand U_n of a user n is represented as Equation 1. The desired positions where user gets on and off are r_n and d_n, and the desired time when user gets on and off are rt_n and dt_n.

$$U_n = (r_n, d_n, rt_n, dt_n) \tag{1}$$

A set of electrical buses B is represented as Equation 2. An element b in Equation 2 is represented as Equation 3, where g is a coordinate of destination, a pair of x and y is a present location, E is a remaining battery charge, and C is a riding capacity. A service area is defined as a 2-dimensional space, and the bus can move freely without receiving the restriction of a road network.

$$B = \{b_1, b_2, \cdots, b_k\} \tag{2}$$

$$b = (g_b, x_b, y_b, E_b, C_b) \tag{3}$$

A set of filling stands for battery charges is represented as Equation 4. An element s is represented as Equation 5, where a pair of x and y is a coordinate. We assume that the battery of an electrical vehicle charges up to 50% of its capacity in 30 minutes.

$$S = \{s_1, s_2, \cdots, s_p\} \tag{4}$$

$$s = (x_s, y_s) \tag{5}$$

A transport plan R, which is represented as Equation 6, means a running route of a bus and includes demands for l users (n=1,. . .,l). An element r is represented as Equations 7, which indicates a user's position (r_r or r_d) or a charge stand position (r_s). An element r is composed of four values, where a pair of x and y is a coordinate, t is an estimated arrival time, and e is an estimated remaining battery charge.

$$R = \{r_1, r_2, \cdots, r_l\} \tag{6}$$

$$r = (x_r, y_r, t_r, e_r) \tag{7}$$

3 Optimization of Transport Schedule for Electrical Vehicles

The genetic algorithm is a major method for VRP, and a gene type is a key issue for the performance of the method. For example, Path Representation (PR) and Genetic Vehicle Representation (GVR) which indicate a solution of VRP are used in [6,7]. In this paper, we apply Path Representation (PR) to our problem.

3.1 Flow of Optimization Process

It is necessary to consider the battery charge timing for on-demand bus system using electric vehicles. Therefore, we apply NIA and GA which are adopted in [3] into our system. NIA schedules transport jobs serially, and the GA minimizes the processing time of the transport jobs in consideration of the battery charging timing. Figure 1 shows the flow of optimization process. The battery power is checked at regular intervals. If the battery power is not enough, a charge stand position is inserted in the transport route by NIA. Until the next battery check, user's getting on/off positions are inserted in the transport route by NIA. Moreover, after scheduling by NIA, GA also applied to the transport route to minimize traveling distance.

3.2 Node Insert Algorithm (NIA)

The detail of NIA for electrical vehicles is shown as below. There are two schedule patterns by NIA: insertion of charge stand positions and user's positions.

Insertion of Charge Stand Position. The nearest charge stand is selected from the set of stands and inserted in the running route R by NIA, when the battery power is less than a constant value. The inserted positions in R must satisfy with two conditions. The first condition is the time limit of all assigned users to the bus, and the second is the battery limit to process all user's demands.

Insertion of User's Position. When a new demand is generated, user's getting on position and off position (r_r, r_d) are assigned to a bus by the system. An insertion position r_d must exist from r_r in the rear side. After the insertion of user's position, the route R is checked in the same way of the insertion of charge stand.

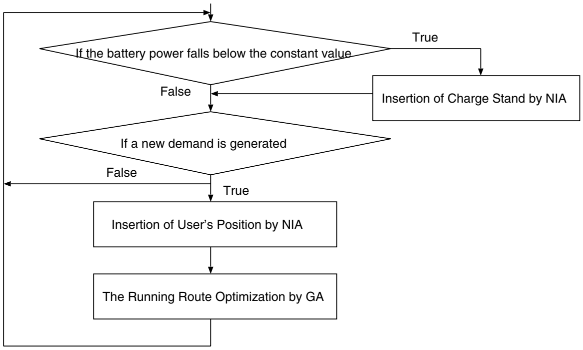

If the battery power falls below the constant value
True
False
Insertion of Charge Stand by NIA
If a new demand is generated
False
True
Insertion of User's Position by NIA
The Running Route Optimization by GA

Fig. 1. Flow of Optimization Process

3.3 Genetic Algorithm (GA)

Here, we explain an optimization process by GA. A sequence of genes for GA represents a set of transport routes, and the type of gene is either getting on/off position r_r/r_d or stand position r_s. Moreover, we describe a population at the generation time T as $P(T)$ in evolution process. First, all patterns of the transport routes by NIA are used as an initial population $P(0)$. Second, two transport routes are randomly selected from the initial population, and two genetic operations (mutation and crossover) are applied to the transport routes. However, genes which are applied mutation and crossover operations may be lethal genes (i.e., genes cannot be solutions of the problem) Here, we define lethal genes which satisfy each of following three conditions.

1. The number of the same type of genes in the sequence is mismatched before and after crossover operation.
2. The getting on position r_n is in the rear of the getting off position d_n in the sequence.
3. If the sequence includes stand positions, the remaining battery runs out until reaching the last position in the sequence.

If a lethal gene is generated after adapting two operations, the lethal genes are restored to keep low convergence time of GA. A lethal gene caused by the two operations must be repaired to keep convergence time low. A repair process of the lethal genes is explained later. Finally, the routes obtained by the operations are ranked in descending order by a fitness function, and high rank routes are selected and added to the next generation $P(T+1)$. These operations are repeated continuously. If T exceeds the maximum generation, the optimization process is finished. As a result, the best sequence of genes in the last generation is selected as a transport route for a bus.

Fitness Function. Here, we define a fitness function for GA as Equation 8. This function represents the weighted sum of the objective functions as Equation

9 and Equation 10. In the equations, $\hat{d}$ represents the total traveling distance by the route, v is an average speed of buses, and E_n is the consumption of battery amount for each unit distance. The first function represents the time required to finish the transportation by the routes, and the second objective function represents the ratio of the remaining battery at the terminal position in the route.

$$f = w_1 \times f_1 + w_2 \times f_2 \tag{8}$$

$$f_1 = \frac{\hat{d}}{v} \tag{9}$$

$$f_2 = 1 - \left(\frac{E - (\hat{d} \times E_n)}{E_{max}} \right) \tag{10}$$

Mutation. An example of mutation process is illustrated in Figure 2. First, two genes are selected from transport route R at random, and the genes (r_{s_1} and r_{d_1}) in route R are mutually exchanged. As a result, a new sequence of genes R' is generated based on the route R.

Crossover. An example of crossover process is illustrated in Figures 3. There are two sequences of genes ($R_1 = r_{r_1} - r_{s_1} - r_{r_2} - r_{d_2} - r_{d_1}$) and ($R_2 = r_{r_1} - r_{r_2} - r_{s_1} - r_{d_1} - r_{d_2}$). If the sequences include no stand positions, a crossover point is selected from the sequence of genes at random. Otherwise, a crossover point is selected to avoid lethal genes in following ways. First, the next stand positions in R_1 and R_2 are selected (r_{r_2} in R_1, r_{d_1} in R_2). Then, the rightmost in the positions (r_{d_1} in R_2) is selected as a crossover point. As a result, new sequences of genes ($R_1' = r_{r_1} - r_{s_1} - r_{r_2} - r_{d_1} - r_{d_2}$) and ($R_2' = r_{r_1} - r_{r_2} - r_{s_1} - r_{d_2} - r_{d_1}$) are generated in Figure 3.

Repair of Lethal Genes. Here, we show the details of repair process for lethal genes in the above-mentioned three cases. Firstly, Figure 4(a) shows the restore process in case 1. We found that gene r_{d_2} is double in the lethal gene. Moreover, the gene r_{r_2} is lost by crossover operation in the lethal gene. Thus, the gene r_{r_2} is inserted into the lethal gene at random position. This operation is repeated until the check process is completed. Secondly, Figure 4(b) shows the restore process in case 2. The getting on position r_{r_2} exists from the getting off position r_{d_2} in the rear side. Thus, gene r_{r_2} is replaced with r_{d_2} to keep consistently. Thirdly,

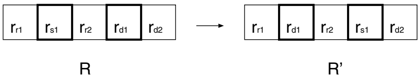


Fig. 2. Mutation Operation

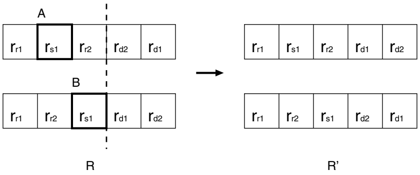
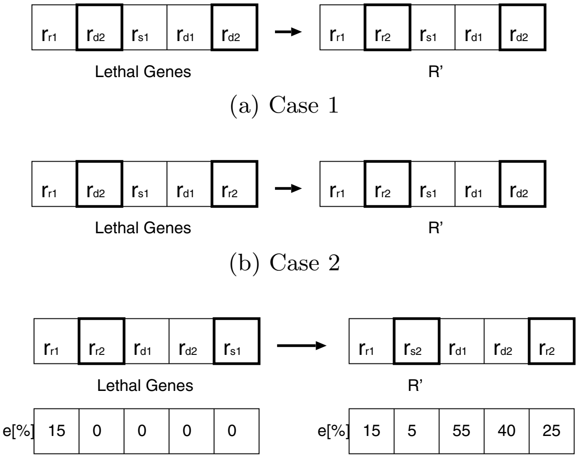


Fig. 3. Crossover Operation



(a) Case 1

(b) Case 2

(c) Case 3

Fig. 4. Repair Process of Lethal Genes

Figure 4(c) shows the restore process in case 3. The stand position r_{s_1} is moved to fore positions in the sequence of the lethal gene to keep the remaining battery $e[\%]$. If another stands is more suitable (i.e., The way to the stand is shorter than other ways), the gene r_{s_1} is replaced with the gene of another stand position.

4 Experiments

4.1 Experimental Setting

Here, we evaluate the effects of our method for on-demand bus using electrical vehicles by computer simulation. The assumption for our simulation is as follows. The service area and time are based on **On-Demand Bus for Nakamura-Machi, Japan**[1]. Time limit of user's demand is randomly set to between 15 and 30 minutes, and electrical vehicles need 30 minutes for half charge of battery. We report two experimental results; the first compares four optimization methods (NIA, NIA+Mutation, NIA+Crossover, and NIA+Mutation+Crossover), and the second compares three weight parameter patterns $w_1 : w_2$ ($5 : 5$, $7 : 3$, and $3 : 7$). The parameter setting is summarized in Table 1.

4.2 Experimental Results

Comparisons of Optimization Methods Figure 5 shows the comparisons of four optimization methods (NIA, NIA+Mutation, NIA+Crossover, and NIA+ GA), and the two bars in the graph represent the acceptance rate and running

[1] Demand-Bus for Nakamura-Machi, Japan is a real transport service, and its URL is http://www.kochi-seinan.co.jp/machif.html

Table 1. Experimental Setting

Parameter	Value	Parameter	Value
Service Area	3 kilometers squared	Vehicle Capacity C	$infinity$
Unit Time	6 minutes	Stand Size p	3
Demand Probability	1 per 6 unit times	Number of Gene	10
Vehicle Size k	2	Max Generation T	15

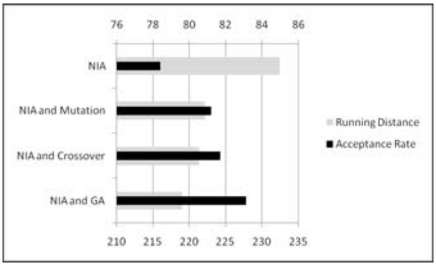


Fig. 5. Comparisons of Optimization Methods

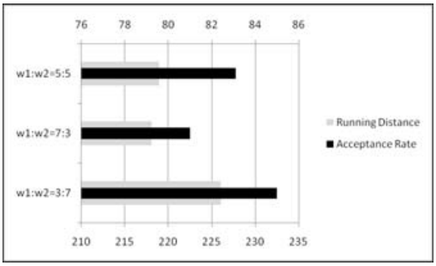


Fig. 6. Comparisons of Weight Parameters

distance of electric vehicles. We found that the mutation and crossover operations improve the efficiency of the system, especially when the both are together applied to the schedule.

Comparisons of Weight Parameters. Figure 6 shows the comparisons of three weight patters (5 : 5, 3 : 7, and 7 : 3). We found that the weight parameter is able to adjust the balance between the running distance and acceptance rate. Thus, we must decide the parameter so that the efficiency of the system maximizes according to the demands.

5 Conclusions

In this paper, we focused on the on-demand system which is more flexible transportation compared to traditional bus systems. It is necessary to consider transport schedules depending on the battery charging so as to introduce electrical vehicles to the system. Therefore, we proposed a scheduling method based on GA in consideration of battery charge timing. As a result, the acceptance rate is improved by 5 %, and it also decreases the running distance. Moreover, we found that it is possible to improve the acceptance rate by always keeping constant battery. The proposed system can reduce the emitter of carbon dioxide by about 60%, the running cost by 80% compared to diesel buses. Consequently, we can say that the proposed system can perform in environment-friendly way. As future work, we intend to introduce a user exchange approach among vehicles at the charge stands by sharing routes among vehicles in to decrease the riding time of users at the charge stands. If the situation occures, the customer's time required increases in vain. To resolve the problem, it is necessary to optimize the charge timing by sharing the route with several vehicles.

Acknowledgment

We gratefully acknowledge the advice and criticism of Prof. Naoyuki Uchida of Tokyo University of Science, Japan.

References

1. Harano, T., Ishikawa, T.: On the validity of cooperated demand bus. Technical Report 2004-ITS-19-18, Technical Report of IPSJ (2004) (in Japanese)
2. Noda, I., Shinoda, K., Ohta, M., Nakashima, H.: Evaluation of usability of dial-a-ride system using simulation. Journal of Information Processing Society of Japan 49(1), 242–252 (2008)
3. Keiichi, U., Ryuji, M.: A real-time dial-a-ride system using dynamic traffic information. The Transactions of the Institute of Electronics, Information and Communication Engineers 88(2), 277–285 (2005)
4. Solomon, M., Desrosiers, J.: Time window constrained routing and scheduling problems. Transportations Science 22, 1–13 (1988)
5. Desrochers, M., Lenstra, J., Savelsbergh, M., Soumis, F.: Vehicle routing with time windows: Optimization and approximation. Vehicle Routing: Methods and Studies, 65–84 (1988)
6. Russel, M.A., Lamont, G.B.: A genetic algorithm for unmanned aerial vehicle routing. In: Proceedings of the 2005 conference on Genetic and evolutionary computation, pp. 1523–1530 (2005)
7. Tavares, J., Machado, P., Pereira, F.B., Costa, E.: A genetic algorithm for unmanned aerial vehicle routing. In: Proceedings of the 2003 ACM symposium on Applied computing, pp. 753–758 (2003)

Public Large Screen Enabled Content Collection and Connection

Kosuke Numa[1], Hironori Tomobe[1], Tatsuo Sugimoto[1], Masako Miyata[2], Kiyoko Toriumi[1], Jun Abe[1], and Koichi Hori[1]

[1] The University of Tokyo, 4–6–1 Komaba, Meguro, Tokyo, Japan
[2] Sapporo Otani University, 1–1 Kita 16, Higashi 9, Higashi, Sapporo, Japan
numa@ai.rcast.u-tokyo.ac.jp

Abstract. In this paper, we propose a framework for content collection and connection enabled by public large screens and mobile phones. Making people express their stories will encourage their active attitudes toward information management. With the proposed framework, we aim to overcome difficulties on managing flooding information. We applied the framework on two practice oriented systems and held workshops using them.

1 Introduction

Nowadays, we people are living in rapidly increasing quantity of information and feeling difficulties in managing them. A new way is required to manage this flooding information in the coming intelligent society.

We consider that one good way to manage information is to publish new information. To publish their own thoughts, people need to collect related information, think over them, create new relations among them, and add their own opinions. A publishing process requires whole activities around information management. Blogging is one example of such practices. It changed people's attitudes toward information on the Web. In this research, we aim to design new relationship between people and information in the real world. To manage everyday information, people need to express their stories and exchange them. Moreover, a way to connect stories, not just to collect stories, is required since people's stories are not separately concluded.

In this paper, we propose a framework for content collection and connection enabled by large screens and mobile phones. We develop two types of practice oriented systems and organize workshops using them. This research does not provide thorough solution to whole design of social interaction with information in the real world; we show a pilot design of relationship between contents and people. In the next section, we describe the background of this research and our framework. Based on the framework, we show two practices in Sections 3 and 4. In Section 5 we provide discussions and we conclude this paper in Section 6.

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 664–671, 2009.

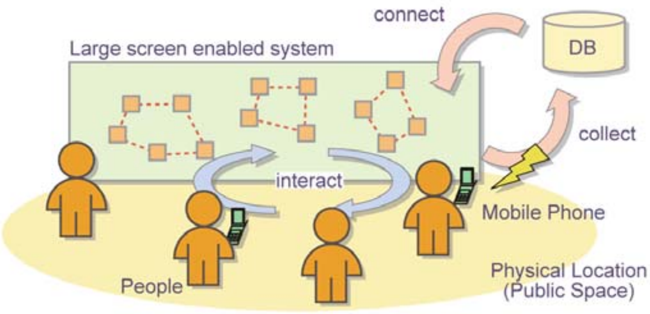


Fig. 1. Proposed framework for content collection and connection

2 Large Screen-Enabled User-Contributed Systems

Storytelling or narrative approaches are getting widely accepted in several fields such as psychology, folklore, education, and therapeutics. It is said that people articulate and interpret their temporal experience by telling stories[7,8]. Storytelling helps people to understand and manage their experiences. First, we aim to make ordinary people tell their stories.

We consider that stories exist not separately but within relations among each other. In hypertext systems like WWW, contents are connected to others. Connected stories are weaved up to larger stories. Second, we aim to connect collected stories.

We employ mobile phones as expressing tools. Recently, mobile phones are usually equipped with cameras. People can take photos and videos and can send them to their friends with their mobile phones. Moreover, some people upload their photos to their blogs or photo sharing sites from their phones. Now, with our mobile phones, we can exchange our experience everywhere in real time — ideally. But most of mobile contents are separately concluded.

We propose a framework to connect and collect people's stories. Figure 1 illustrates outline of the framework. A large screen enabled user contributed system is installed in a public space where people can freely access. People publish their stories using mobile phones. The system stores collected contents and connects them. Connected stories are projected on the screen and people there can interact with them. People's stories form larger stories and return to people. People can change presented connections by adding new contents. In the system, people's stories are iteratively weaved up.

We will find digital signages everywhere in near future[3,2]. They can be shared places for content creation and connection. Peltonen and his group developed an interactive collaboration system using a multi-touch large screen and mobile phones[6]. Their research is technically similar to this research, but we aim at content connections rather than direct interactions among participants.

In this paper, we describe our two workshops which were designed based on this framework. A term workshop here we mean a specially managed place for

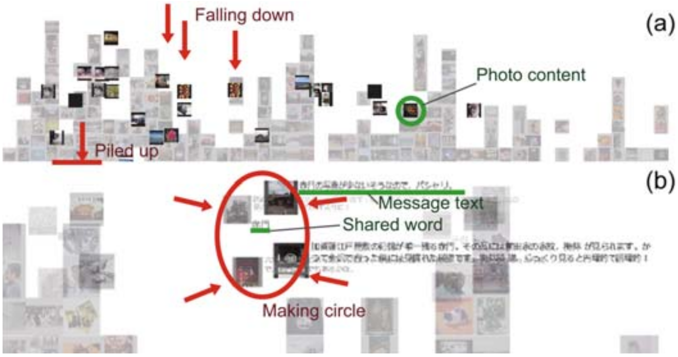


Fig. 2. Screenshots of the shower-style view

practice where people gather and act collaboratively. We regard a workshop itself as a creativity support system[4]. Our final purpose is to redesign relation between people and contents in everyday lives — not in closed spaces like workshops. But the field practices will provide clues to this purpose.

The first practice is the collective photo collage workshop[5]. We collected photo-attached mobile messages and projected them on a screen in a public space. Contents are automatically connected based on their text. The second practice is the mobile video workshop[1]. We collected and presented mobile videos which record interviews connected by question and answer game. In this practice, contents themselves have connections to other contents. We describe the first workshop in Section 3 and the second in Section 4 respectively.

Our purposes in this research are to draw people's stories and to connect them. It is fundamentally important for people to know that their own stories affect whole stories — to know that one is a part of the world. Our practice is designed to simplify and to emphasize this point.

3 Collective Photo Collage Workshop

The first workshop is a pilot practice for collecting and connecting mobile pictures. We developed a system which stores photo-attached messages and automatically connects them. In this section, we introduce this practice briefly.

The workshop was held at a new hall in the University of Tokyo in Japan, which was built on March, 2008. In the workshop, we collected photo-attached e-mails from people around/related to the hall. Collected contents are connected based on shared words included in messages. Messages are morphologically analyzed and shared words are automatically extracted. The system stores contents (photos and text messages) and these data (extracted words) to a database.

Collected contents were shown on two views on site (the slide show view and the shower-style view) and on the Website. The shower-style view connects messages based on shared words. Figure 2 shows screenshots of the shower-style view. Collected contents fall down from the upper edge of the screen and are

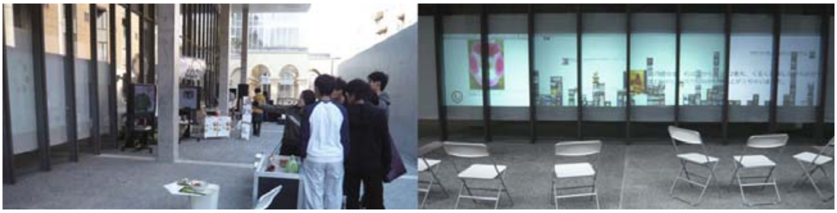


Fig. 3. Scenes from the collective photo collage workshop

piled up on the bottom edge (Figure 2(a)). After all entries are piled up, several entries form a circle based on a randomly selected shared word (Figure 2(b)). Minimum number of entries for each circle is four.

As shown in Figure 3, the slide show view and the shower view are projected onto the screen. Passengers can see entries on the screen, and can post their own entries on site with their mobile phone. The posted entries are soon displayed on the slide show, and will be used for word-based circles.

We began to collect contents online from 22nd February 2008. The workshop was held 21st through 28th March. Through the event, 351 entries are posted by 184 users. Since the detailed analysis is described in [5], here we introduce just several findings. There were roughly two types of participants: participants who posted multiple contents online, and participants who posted single contents for each from the venue. While 88.6% of contents were connected to other contents, participants on site tended to communicate with people/things there rather than to interact with contents on the system. The aim that connected stories form large stories was not fully achieved. A participant also claimed this point – "I want to put my photo directly into a circle."

4 Mobile Video Workshop

Contents were not directly connected in the first workshop; connections were generated automatically. We devised the system to show direct connections.

4.1 Outline

We designed a workshop program called "*Keitai Trail!*" and practiced it in conjunction with Ars Electronica festival 2008. Ars Electronica festival is one of the most popular media art festival held annually in Linz, Austria. A word *Keitai* is a mobile phone in Japanese.

In this workshop, we record people's stories on mobile videos and connected them directly based on a rule. Figure 4 illustrates the workshop scenario (and the system architecture explained in the next subsection). This workshop is designed to be held not only at a single place but also outside space around a main venue.

The main venue is a kind of base where facilitators present a progress and participants' expressions of the workshop. Facilitators go outside and ask people there to join the workshop. If one accepts, facilitators shoot a video of her talk

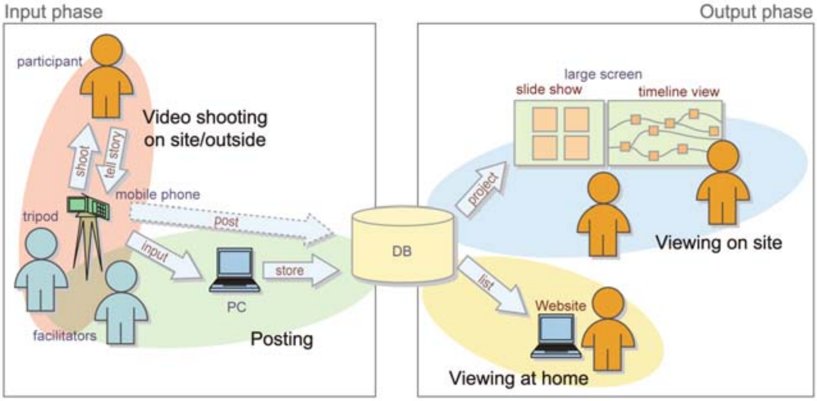


Fig. 4. Workshop scenario and system architecture

with a mobile phone. In the main venue, connected videos are shown on large screens by an installed support system which we describe in the following section.

Participants' talks are requested to follow the "talking format" shown in Figure 5. The format connects talks in a simple rule. The format consists of four parts: (1) an answer to a question from a former participant, (2) a short free talk, (3) a connecting phrase to the next part, and (4) a question to a next participant. The question in the fourth part will be answered in a next participant's first part. This is like a question and answer game. A question from a former participant is a cue to a free talk; a talk in the second part is a core of her story. Participants can choose a question to answer, i.e., her former participant. With this rule, a story is connected to other stories. Connected stories make large stories. This format derives people's stories and connects them. Figure 6 shows example stories and their connections.

4.2 System

Figure 4 illustrates a usage scenario of the installed system which consists of two phases.

In input phase, facilitators shoot a video of a participant telling story with a tripod equipped mobile phone. Facilitators store video files and their connections to a database. It is theoretically possible to post videos directly from mobile phones. But in this case, we needed to develop an input interface for PCs due to temporal technological limitations such as maximum size of uploading files and covered service in roaming area of mobile phones we used.

In the output phase, we prepared two ways of viewing. At the main venue, two types of interfaces are projected on large screens. Slide show view plays recently posted two and randomly selected two videos at a same time. Timeline view is designed to show whole connections of videos (Figure 7). In the view, nodes represent videos and arcs represent connections. The x-axis direction stands for time and the view can be scrolled in this direction. A participant can trace whole

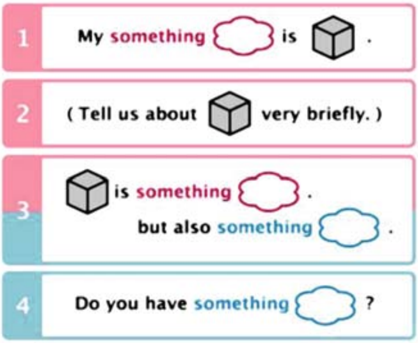


Fig. 5. Talking format we used in our practice

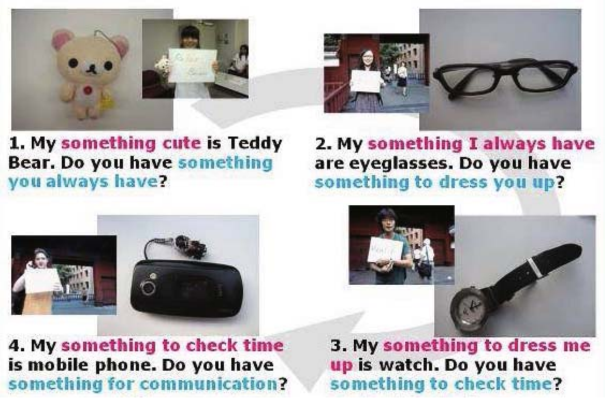

1. My something cute is Teddy Bear. Do you have something you always have?
2. My something I always have are eyeglasses. Do you have something to dress you up?
4. My something to check time is mobile phone. Do you have something for communication?
3. My something to dress me up is watch. Do you have something to check time?

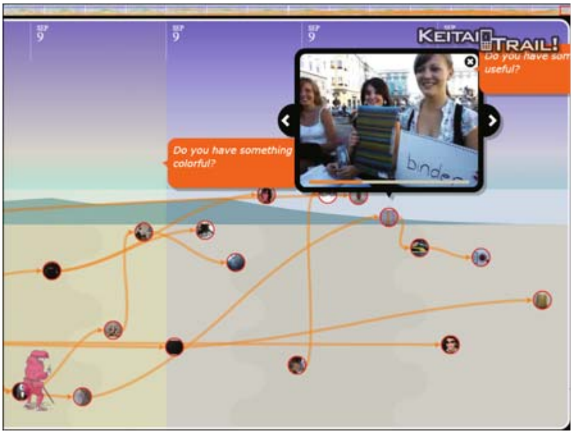
Fig. 6. Example stories and their connections

stories and can add her story. Participants who joined outside the venue have two options to see their own videos. They can visit the venue of course. In addition, we provide a Website which lists collected videos. They can browse stories at home.

4.3 Result

We held the workshop during 4th to 9th September 2008 in Linz, Austria. Finally, we collected 258 stories (videos): 218 in this workshop and 40 in the preliminary workshop held in Japan. Scenes from the workshop are shown in Figure 8

These videos are divided into five clusters. The largest cluster includes 174 videos, and has 28 branches. The longest path in the cluster is 38 connections (39 videos) in length. Most of stories are connected to this cluster or the second largest cluster including 77 videos. The other three clusters were cut off from the larger clusters by a few participants who ignored former contents. Most of stories occupy parts of large stories.



Do you have som useful?

Do you have something colorful?

Fig. 7. Screenshot of timeline view

Fig. 8. Scenes from the mobile video workshop (the venue and outside)

5 Discussion

Through two workshops, we connected contents. In the first workshop, connections were generated automatically. In the second workshop, we designed the system to connect contents directly and manually. As a result, we obtained more active commitment from participants in the latter workshop.

These two workshops do not differ so much technically. Contents are stored to databases, the contents and their connections are output in XML data. Interfaces read the data and present them in designed views. Interactions in the workshops, however, were different. While a content and its relation could not be seen just after it was posted in the former workshop, we showed direct connections of contents in the latter workshop. One of the reasons for the difference of the results was this difference of interaction design.

6 Conclusion

In this paper, we proposed a framework for content collection and connection enabled by public large screens and mobile phones. We applied the framework on two practice oriented systems and held workshops using them. Through this paper, we wanted to claim that expressing a single story is a small activity, but it has a big meaning socially.

Acknowledgements

This work has been supported by a grant from the Japan Science & Technology Agency under CREST Project.

References

1. Abe, J., Toriumi, K.: Collaborative Narratives in the Digital Age: An Analysis of “Keitai Trail! - Mobile Video Workshop -”. In: Proceedings of Annual Workshop on Digital Communication (2009)
2. Kern, D., Harding, M., Storz, O., Davis, N., Schmidt, A.: Shaping How Advertisers See Me: User Views on Implicit and Explicit Profile Capture. In: CHI 2008 Extended Abstracts (2008)
3. Müller, J., Schlottmann, A., Krüger, A.: Self-optimizing Digital Signage Advertising. In: Adjunct Proceedings of Ubicomp 2007 (2007)
4. Numa, K., Toriumi, K., Tanaka, K., Akaishi, M., Hori, K.: Participatory Workshop as a Creativity Support System. In: Lovrek, I., Howlett, R.J., Jain, L.C. (eds.) KES 2008, Part II. LNCS, vol. 5178, pp. 823–830. Springer, Heidelberg (2008)
5. Numa, K., Tomobe, H., Tanaka, K., Nishimura, T., Hori, K., Sunaga, T.: A Case Study on Interactions with User Contributed Website in Public Space. In: The 7th International Workshop on Social Intelligence Design, SID 2008 (2008)
6. Peltonen, P., Salovaara, A., Jacucci, G., Ilmonen, T., Ardito, C., Saarikko, P., Batra, V.: Extending Large-scale Event Participation with Usercreated Mobile Media on a Public Display. In: Proceedings of the 6th International Conference on Mobile and Ubiquitous Multimedia (2007)
7. Ricpœur, P.: Temps et Récit. Seuil (Time and Narrative. trans. Kathleen McLaughlin and David Pellauer. University of Chicago Press (1984)
8. Schank, R.C.: Tell Me a Story: Narrative and Intelligence. Northwestern University Press (1990)

Implementing Multi-relational Mining with Relational Database Systems

Nobuhiro Inuzuka* and Toshiyuki Makino

Nagoya Institute of Technology,
Gokiso-cho Showa, Nagoya 466-8555, Japan
inuzuka@nitech.ac.jp, makino@nous.nitech.ac.jp

Abstract. Multi-relational data mining (MRDM) is to enumerate frequently appeared patterns in data, the patterns which are appeared not only in a relational table but over a collection of tables. Although a database usually consists of many relational tables, most of data mining approaches treat patterns only on a table. An approach based on ILP (inductive logic programming) is a promising approach and it treats patterns on many tables. Pattern miners based on the ILP approach produce expressive patterns and are wide-applicative but computationally expensive because the miners search among large pattern space. We have been proposing a mining algorithm called MAPIX[3]. MAPIX has an advantage that it constructs patterns by combining atomic properties extracted from sampled examples. By restricting patterns into combinations of the atomic properties it gained efficiency compared with conventional algorithms including WARMR[1,2]. In order to scale MAPIX to treat large dataset on standard relational database systems, this paper studies implementation issues.

1 Introduction

Relational pattern mining has been disscussed in the framework of multi-relational data mining (MRDM) and it is suitable to use the technique of inductive logic programming (ILP). Warmr[1,2] is a representative algorithm along this context.

Warmr generates candidate patterns in a top-down way from simple to complex in level-wise. It stops grow a pattern more complex once it finds the pattern infrequent. It is similar to the principle used in Apriori[5]. In spite of the cut-down procedure it has limitation, because of the exponentially growing space of patterns with respect to the length of patterns and the number of relations. Mapix acquired much efficiency at the sacrifice of the variety of patterns. It only finds patterns as combination of attributes, which are dynamically constructed as a set of first-order literals from given examples. It is bottom-up in the sense that attributes are not given in advance but are constructed from given examples. It first constructs all attributes, called property items, which are appeared

* This reserach is partially supported by JSPS, Grant-in-Aid for Scientific Research (C) (20500132).

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 672–680, 2009.

train
t_1
t_2
t_3
t_4
t_5

has-car	
t_1	c_1
t_1	c_2
t_2	c_3
t_3	c_4
t_3	c_5
t_4	c_6
t_4	c_7
t_5	c_8

triangle
c_1
c_2
c_4
c_7

circle
c_2
c_5
c_6
c_8

Fig. 1. The database $\mathrm{R}_{\mathsf{train}}$ which has four tables including key table train

in examples. Then it applies an Apriori-like procedure for the property items. It succeeded to prohibit duplication of patterns in the sense of logical equivalence.

In this paper we study on implementation of Mapix for relational database management system (RDMS). Unsual setting of MRDM bases implemation on Prolog system and data are assumed to be manimulated on main memory. In order to apply MRDM methods for wider application areas, we try to implement Mapix on RDMS using SQL manipulation.

2 Multi-relational Pattern Mining

Using a simple example we introduce multi-relational pattern mining and Mapix algorithm for it. Consider a databalse $\mathrm{R}_{\mathsf{train}}$ including four relational table as shown in Fig. 1. A relation $\mathsf{train}(\cdot)$ keeps trains and $\mathsf{has\text{-}car}(\cdot,\cdot)$ shows cars to which each train connects. Other two tables $\mathsf{triangle}(\cdot)$ and $\mathsf{circle}(\cdot)$ show attributes of loads kept in cars. For multi-relational pattern mining we choose a table (a *key* in Warmr's term) in the database and try to find patterns which are appeared in many objects, more objects than a prescribed threshold, in the chosen table. Such patterns are called frequent patterns. For example we may see a pattern that many train has at least two cars of which a car keeps a triangle shaped load and the other keeps a circle shaped load. This pattern can be described by the following relational formula.

$$\mathsf{train}(A) \wedge \mathsf{has\text{-}car}(A, B) \wedge \mathsf{has\text{-}car}(A, C) \wedge \mathsf{triangle}(B) \wedge \mathsf{circle}(C).$$

A conventional successful algorithm Warmr finds frequent patterns in a top-down way. That is, it generates and tests patterns from simple to complex. If a simple pattern is found infrequent Warmr does not try the pattern grow longer. Mapix has a different strategy for finding patterns. It restricts patterns into combination of basic patterns, called property items. Property items can be seen as a natural extension of attributes in first order logic and we see it in later paragraphs. Mapix also restricts only property items appeared in sampled objects. By the restriction Mapix does not have completeness, that is, it does not enumerate all frequent patterns. Mapix gained much efficiency at the sacrifice of the restriction, although the successor of Mapix has closed to complete enumaration[4].

The reason of inefficiency of WARMR is not only by the top-down method. It comes from duplication of patterns. Two differently appeared pattern may be equivalent logically and it is difficult to cut down all equivalent patterns. MAPIX avoids all logical duplication in searching patterns[3,4].

2.1 Property Items

We assume readers familiar with terms of logic programming. Arguments of predicates are given execution mode. Input mode is denoted by + and output is denoted by −. We assume modes of the predicates in $\mathrm{R}_{\mathsf{train}}$ as $\mathsf{has\text{-}car}(+,-)$, $\mathsf{triangle}(+)$ and $\mathsf{circle}(+)$. We give output modes to all arguments of key predicate in a technical reason.

In MAPIX, predicates are classified into two types. Predicates which have only input mode arguments are called check predicates. Predicates which include output mode arguments are called path predicates. $\mathsf{has\text{-}car}(+,-)$ is a path predicate and $\mathsf{triangle}(+)$ and $\mathsf{circle}(+)$ are check predicates. We call a literal of check (path) predicates a check (path) literal.

A path literal has a function like a mapping, a path literal derives a term as an output from an input term. A check literal has a function like an attribute. It takes some terms and describes a character, such as its shape. For example, for a train t_1, that is a literal $\mathsf{train}(t_1)$ is recorded in the key table, let us imagine a set of literals,

$$\{\mathsf{has\text{-}car}(t_1, c_1), \mathsf{triangle}(c_1)\}.$$

The first literal $\mathsf{has\text{-}car}(t_1, c_1)$ has a function deriving a car c_1 from t_1 and then the second literal describes an attribute for c_1 as it is a triangle.

The set of literals can be seen an extended attribute. An extended attribute has two parts. One consists of path literals and they derive from a term to a term and the other part consists of a check literal and describes a fact referring the derived terms. We call such an extended attribute a property item.

A property item is generalized by replacing terms by variables and represented as follows.

$$\mathsf{train}(A) \leftarrow \mathsf{has\text{-}car}(A, B) \wedge \mathsf{triangle}(B).$$

In this formulation, we used a clausal formula for a property item, where a key literal is given as a head and path and check literals are combined by conjunction in its body part. For detailed terminology and semantics of formulae are given in [3].

2.2 MAPIX Algorithm

An outline of MAPIX algorithm is as follows.

1. It samples a set of examples (tuples) from a key table.
2. For each example it collects all relevant literals from database.
3. For the set of relevant literals of each example it extracts and generates all property items.

4. Extracted property items are gathered together and logically duplicated property items are eliminated.
5. It measures frequencies of property items in databases and eliminates infrequent property items.
6. By using an analogous method to Apriori[5] it enumerates all frequent patterns made by conjunction of property items.

Relevant literals of an example are literals in databases which has connection to the example. That is it can be defined as follows.

- The example itself is an relevant literal.
- If every input mode argument of a literals in the database is appeared in an output mode arguments of relevant literal, the literal is also relevant literal.

The relevant literals keep all information of the example. For example relevant literals of the example $\mathsf{train}(t_1)$ are,

$$\mathsf{train}(t_1), \mathsf{has\text{-}car}(t_1, c_1), \mathsf{has\text{-}car}(t_1, c_2), \mathsf{triangle}(c_1), \mathsf{circle}(c_2).$$

The idea of relevant literal is related to the idea of saturation clauses in ILP literature[6,7].

3 An Implementation Combining Relational Database Management Systems

Pattern mining algorithms based in ILP approach usually treat data on main memory but not on database in storage system. Such algorithms usually assume an environment of logic programming, such as Prolog system. The environment has advantages that logical manipulations are easier and extracted logical patterns can be combined with other knowledge-base immediately. However, usability of such algorithms is limited because they require transformation of data into logic programming environment and large scale data can not be manageable.

In this section we give an implementation of MAPIX algorithm combining relational database management systems.

3.1 SqlMapix

We assume all data are kept in a database in DBMS. Our implementation uses Prolog and ODBC (open database connectivity) interface in order to manipulate database from a Prolog program. We also assume that the Prolog program keeps database scheme and input/output-mode information of data tables. We also give the Prolog program a name of key table and a frequency threshold. That is our Mapix system, which we call SqlMapix, is summarised as follows.

Input in DBMS: A database.

Input of SqlMapix **in Prolog:** A frequency threshold, schemes of tables in the database in DBMS, attribute types and input/output modes of the tables, and a name of key table.

Output of SqlMapix**:** All frequent patterns appeared in DBMS the pattern which are derived by combining property items.

Here we give a Mapix implementation using Prolog system with access to database on DBMS. Considering that database in DBMS can be processed on DBMS, we carefully divide manipulations of Mapix algorithm into ones in a Prolog program and ones on DBMS. For dividing manipulation we assume that a set of relevant literals of an example is enough small to manipulate on main memory and assume that any relation table may be too large to bring on main memory. Hence we strictly prohibit to take manipulations over whole data of a table.

We have the following procedure by these consideration. Lines headed by **DBMS** is manipulation on DBMS and ones headed by **Prolog** is manipulation by a Prolog program.

DBMS. It samples examples (tuples) from a key table. The sampled examples are stored in a table on DBMS.

DBMS. It generates relevant literals of sampled examples by a SQL manipulation. All relevant literals of examples are stored in a table.

Prolog. For each sampled example the set of relevant literals of the example are transmitted to a Prolog process and transformed into atoms in logical formula.

Prolog. It processes the relevant literals of an example in logical formulae and generates property items.

Prolog. All property items are gathered and logical duplications are eliminated.

Prolog. For each property item it generates SQL query to test if each example satisfies the property items. The queries are issued for DBMS.

DBMS. A transaction table is generated. In the transaction table property items that an example satisfies are recorded. It is similar to a market basket database in which items bought by each custom are recorded.

DBMS. By processing using Apriori-like procedure, it generates all frequent combination of property items in the transaction database.

That is, operations to sample examples, to generate relevant literals, to generate a transaction database, and to process the Apriori-like procedure, are operated on DBMS.

The following paragraphs explain the each step of DBMS operations.

Sampling Examples from Key Table. Sampling examples is operated on DBMS and it can be done by a standard selection operation using random ordering.

Generating Relevant Literals. Relevant literals are generated by taking join operations between the sampled key table and other relation tables. Columns of the key table can be connected to +-mode columns of other tables if the columns have the same type. If the connected table has −-mode column the column can be connected to other table again. Join operation produces new values for columns

Table 1. A procedure to generate relevant literals

```
generate relevant literals
input   the set $B$ of table names with their schemes, column types and modes;
        the key table $E$ in $B$ with its scheme and column types;
output  the table $S$ of relevant literals;
Let $S \subseteq E$ the table of sampled examples from key table $E$;
Open := the set of all columns of $E$;  % all columns are −-mode.
While Open $\neq \emptyset$ do
    Choose a column $c$ from Open;
    For each table $R$ in $B$ and each +-mode column $d$ of $R$ do
        If $c$ and $d$ are the same type then
        $S := S$ LEFT OUTER JOIN $r$ ON $S.c = R.d$;
        If some raws are joined from $r$ then
            Add all −-mode column of $R$ to Open;
    Open := Open − $\{c\}$;
```

train
t_1
t_2
$\cdots$
$\cdots$

has-car	
t_1	c_1
t_1	c_2
t_2	c_3
$\cdots$	$\cdots$

triangle
c_1
c_3
$\cdots$
$\cdots$

circle
c_2
$\cdots$
$\cdots$
$\cdots$

train	has-car		triangle	circle
t_1	t_1	c_1	c_1	null
t_2	t_1	c_2	null	c2
$\cdots$	t_2	c_3	c_3	null
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$

Fig. 2. An example of generating relevant literals by join operations. The original tables (left) and the resulted tables (right).

introduced by the joined table when the connected columns have the same value. Natural join operation X **JOIN** Y **ON** $X.c = Y.d$ is an operation described:

$$X \textbf{ JOIN } Y \textbf{ ON } X.x_i = Y.y_j$$
$$= \{(x_1, \ldots, x_m, y_1, \ldots, y_n) | (x_1, \ldots, x_m) \in X, (y_1, \ldots, y_n) \in Y \text{ and } x_i = y_j\}$$

It eliminates raws of X when it does not match with any raws of Y. For our purpose we need remain raws which does not match with any raws of the other table. We use **LEFT OUTER JOIN** for this purpose. It keeps all raws of X even if it does not match with Y. When $(x_1, \ldots, x_m) \in X$ does not match with any raws of Y, the operation keeps the raw with null values for columns for unmatched columns, i.e. $(x_1, \ldots, x_m, \text{null}, \ldots, \text{null})$ in result.

$$X \textbf{ JOIN } Y \textbf{ ON } X.x_i = Y.y_j$$
$$= \{(x_1, \ldots, x_m, y_1, \ldots, y_n) | (x_1, \ldots, x_m) \in X \wedge (y_1, \ldots, y_n) \in Y \wedge x_i = y_j\}$$
$$\cup \{(x_1, \ldots, x_m, \text{null}, \ldots, \text{null}) | (x_1, \ldots, x_m) \in X \wedge \neg\exists (y_1, \ldots, y_n) \in Y . x_i = y_j\}$$

When some raws get new values the operation continues for −-mode columns introduced. This operation has to continue until no new −-columns. The procedure is given in Table 1.

Table 2. A procedure to generate transaction tables for property items

```
generate a transaction table
input  the set $P$ of property items;
output the transaction table $T$;
Let $T$ an empty table;
For each property item $p \in P$ do
        Generate query $q$ to test whether every tuple of the key table $E$ satisfies $p$;
        Execute $q$ and get table $T_p = \{e \in E | e \text{ matches } withp\}$;
        Add a column keeping the ID number of $p$ to $T_p$,
                i.e. $T_p := \{(e, i) | e \in T_p$ and $i$ is the ID of $p\}$;
        Union $T_p$ to $T$;
```

train	property item ID
t_1	i10
t_1	i12
t_2	i10
t_2	i11
...	...

Fig. 3. Transaction database that shows property items satisfied by each example

The SQL queries for the join operation is generated by a Prolog program. In order to produces queries the Prolog program need keep the information of schemes of tables, and types and modes of every columns of tables.

An example of generating relevant literals by the procedure is shown in Fig. 2

From the resulted table of relevant literals, lines for each examples are transmitted to the prolog program. The lines are transformed into logical formulae. Then property items are extracted from the relevant literals. We used the procedure given in [3] for this purpose.

Generating a Transaction Table. Transaction tables can be generated by the procedure in Table 2.

Enumerating Frequent Patterns. From transaction database frequent patterns combining property items are enumerated by Apriori-like method. Implementation on DBMS is investigated in [8].

SqlMapix has limitation in the predicate mode. In this version of implementation it does not allows path predictes with more than one input mode arguments. There is another limitation which comes from DBMS specification. A table for relevant literals can take a large number of columns and the number is restricted by DBMS. These limitation shuld be removed in future work.

Examples sampled	runtime (sec.)	
for extracting pr. items	Mapix	SqlMapix
100	715	156
1000	7,036	678
3369	23,766	2,792

Fig. 4. Average runtime of Mapixand SqlMapix.

4 Experiments

We examined SqlMapix system using a dataset on grammar structure of English sentences. The dataset includes information of 3369 English sentences from Wall Street Journal. The dataset were prepared in [3]. Sentences are analysed and a relation table has-a-part-of$(+, -)$ and tables for POS tags. has-a-part-of keeps substructures of sentences as has-car is.

Using this dataset we examined to generate a transaction database and measured runtime when 100, 1000 or 3369 examples are sampled for extracting property items. Note that sampled examples are used for extraction of property items but a transaction database is generated for all examples. Table 4 is the result of the experiment. Every runtime is the average of 10 trials. The time were compared with the original implementation of Mapix. SqlMapix processes faster than the original Mapix.

5 Conclusions

We described an implementation of Multi-relational pattern mining algorithm combining DBMS. A difficult point is to combine information across many tables on database. We proposed a method to extract combined information using SQL operations transmitted from a Prolog program. Multi-Relational data mining is powerful but costly in general. The proposing method showed a direction that Multi-relational mining method can be applied with large scale databases. The implementation does not cover general form of dataset. Our future work include to remove the limitation and also to implement the successor algorithm of Mapix for more comprehensive data mining.

References

1. Dehaspe, L., De Raedt, L.: Mining Association Rules in Multiple Relations. In: Džeroski, S., Lavrač, N. (eds.) ILP 1997. LNCS (LNAI), vol. 1297, pp. 125–132. Springer, Heidelberg (1997)
2. Dehaspe, L., Toivonen, H.: Discovery of Relational Association Rules. Relational Data Mining, pp.189–212 (2001)
3. Motoyama, J., Urazawa, S., Nakano, T., Inuzuka, N.: A Mining Algorithm Using Property Items Extracted from Sampled Examples. In: Muggleton, S.H., Otero, R., Tamaddoni-Nezhad, A. (eds.) ILP 2006. LNCS (LNAI), vol. 4455, pp. 335–350. Springer, Heidelberg (2007)

4. Inuzuka, N., Motoyama, J., Urazawa, S., Nakano, T.: Relational pattern mining based on equivalent classes of properties extracted from samples. In: Washio, T., Suzuki, E., Ting, K.M., Inokuchi, A. (eds.) PAKDD 2008. LNCS (LNAI), vol. 5012, pp. 582–591. Springer, Heidelberg (2008)
5. Agrawal, R., Srikant, R.: Fast Algorithms for Mining Association Rules in Large Databases. In: VLDB, pp. 487–499 (1994)
6. Rouveirol, C.: Extensions of Inversion of Resolution Applied to Theory Completion. In: Inductive Logic Programming, pp. 63–92. Academic Press, London (1992)
7. Idestam-Almquist, P.: Efficient Induction of Recursive Definitions by Structural Analysis of Saturations. In: Advances in ILP, pp. 192–205. IOS Press, Ohmsha (1996)
8. Sarawagi, S., Thomas, S., Agrawal, R.: Integrating Association Rule Mining with Relational Database Systems: Alternatives and Implications. In: SIGMOD, pp. 343–354 (1998)

A Simple Method for 3-Dimensional Photorealistic Facial Modeling and Consideration the Reconstructing Error

Ippei Torii, Yousuke Okada, Masayuki Mizutani, and Naohiro Ishii

Aichi Institute of Technology
Management and Information Science
Yachikusa 1247, Yakusa-cho, Toyota-shi, Aichi, Japan
{mac,x07232xx,x09120xx,ishii}@aitech.ac.jp
http://www.ait.ac.jp/

Abstract. The process of creating photorealistic 3-dimensional computer graphic (3DCG) images is divided into two stages, i.e., modeling and rendering. Automatic rendering has gained popularity, and photorealistic rendering is generally used to render different types of images. However, professional artists still model characters manually. Moreover, not many progresses have been achieved with regard to 3-D shape data acquisition techniques that can be applied to facial modeling; this is an important problem hampering the progress of 3DCG. Generally, a laser and a highly accurate camera are used to acquire 3-D shape data. However, this technique is time-consuming and expensive. Further, the eyes may be damaged during measurements by this method. In order to solve these problems, we have proposed a simple method for 3-D shape data acquisition using a projector and a web camera. This method is economical, simple, and less time-consuming than conventional techniques. In this paper, we describe the setup of the projector and web camera, shape data acquisition process, image processing, and generation of a photorealistic image. We evaluate the error margin. We also verify the accuracy of this method by comparing the photograph of a face with its rendered image. After that, we pick up only labial and mouth part from obtained facial modeling data and expand it into animation.

Keywords: Photorealistic 3DCG, Facial Modeling, Shape Reconstructing.

1 Introduction

Recent developments in 3-dimensional computer graphics (3DCG) have made it possible to generate photorealistic images. Research is being conducted on various 3DCG applications and generation of photorealistic images is one of the main research topics. The generation of photorealistic images is divided into two stages, namely, modeling and rendering. Modeling is a process which defines and creates the data of facial shape and rendering is a process whereby the final image is generated from the modeling data. The rendering is comparatively easy for

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 681–688, 2009.

automatic creation and it can generate photorealistic images from complicated form easily. However, modeling must be carried out manually by an artist, and its efficient improvement is mall. Achieving highly precise geometric modeling is an important challenge in 3DCG[1]. Conventional methods such as stereo imaging and 3-D scanning are used for 3-D modeling. Stereo imaging makes use of a stereo camera that can simulate binocular vision, and 3-D scanning involves the use of a laser scanner and a CCD camera. In 3-D scanning, the vertical planes of an object are measured using the laser scanner by the 2-D scanning method, in which a method of measuring the object shape of carrying the cross section is developed. However, conventional methods have some disadvantages. Stereo imaging is difficult to use under varying light conditions because the shadow of the object is misinterpreted by the turbulence light. In 3-D scanning, we can acquire large high-density data by only one scan in short time; however, a large amount of memory is required because it is necessary to scan the object in all directions. In 3-D scanning, the object size in the passage direction is misinterpreted when the speed and direction of objective change in the measurement domain[6]. In this paper, we propose a new method that reproduces the shape data of a human face on the computer in a short period of time. In comparison with the laser scan method, the proposed method is economical, simple, and less time-consuming[10,5].

2 Facial Modeling

It can be confirmed that the line is horizontal even if it is observed from anywhere when the horizontal line projected to the plane. However, the line has been often curved when this line is projected to complex ups and downs, and observed from the upper part. The shape of a face can be acquired by comparing and analyzing this difference with a web camera.

3 Stages in Facial Modeling

3.1 Environment

The basic devices required for facial scanning in this method are a projector and a web camera. The projector and web camera must be set up in the position as shown in Fig. 1. The projector and web camera should be set up based on the lens.

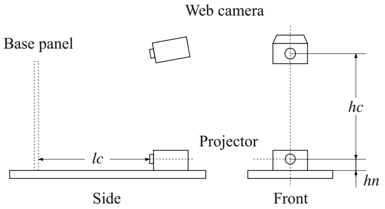


Fig. 1. Projector and web camera set up

3.2 Initialization of Base Line and Base Panel

Before facial scanning, we generated a line of 1 pixel (base line) that is standard on the computer. The line is projected in three colors on the base panel (e.x. white paper), i.e., red(R: 255), green(G: 255), and blue(B: 255), and compared them on the basis of extraction accuracy. We find that red is not suitable for extraction because it interferes with the color of the lips and skin. Green is also inappropriate because it would mask the color of a blood vessel if it is prominent. Finally, blue is used as the background for chroma key as it has the highest extraction accuracy. The base line generated by the computer is projected on the base panel. We also measure the distances between the projector's lens and the base panel (*lc*), the center of the projector lens and web camera (*hc*), and the installation side and center of the projector lens (*hn*). Then, we obtain a static image with the web camera and map the base line on the global coordinate system to the camera coordinate system. The distortion of the camera lens must be corrected.

3.3 Generation and Projection of Scan Line

We generate a scan line projected to the face with the projector based on information of the base line set in the preceding section. The color of the scan line is the same as that of the base line, i.e., blue. The direction of movement of the scan line should be the same as that of the base line. Here, sequential scanning is used in order to reduce processing complexity and mutual interference by multiple base lines. It is necessary for the web camera to take a picture after every line is scanned. This operation is performed at very high speeds, which reduces eye strain. The danger of the scan line passing over the eyes is not as great as that posed by a laser beam; however, the examine must close his/her eyes when a picture is being taken. Only the scan line is extracted from the image data and all additional information is discarded(Fig. 2). The image by web camera is converted RGB color space into HSV (Hue, Chrome, Brightness) color space. It analyzes hue angle of all pixel and defines average angle as threshold H_t. Hue angle H_{ij} from each pixel is got from obtained graphics data. If H_{ij} satisfied $360 - H_t < H_{ij}$ *and* $H_{ij} < H_t$ $(0 < H_{ij} < 360)$, the pixel of H_{ij} is True, while Not is False in Fig. 3. Also True is white, and False is black. Furthermore, the noise in the direction of the x-axis can be removed that deviated from the scan line greatly using continuing on an image. Line thinning is performed in order to increase the clarity of the extracted line. The thickness of the line is set to 1 pixel by deducing the average of the point of the topmost part and the lowermost part of the line.

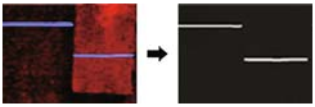


Fig. 2. Analysis of scan line

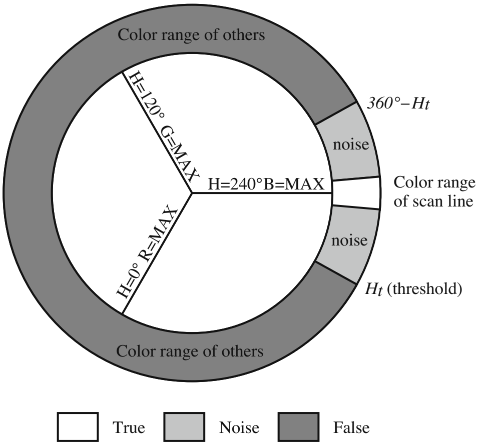


Fig. 3. HSV color space from RGB

3.4 Computation of Coordinate Values

The scanned data must be sampled along the x-axis, and the amount of data must be determined. Minute facial contours are considered as noise. According to the sampling theorem[9], the x-axis should be divided into at least 5 intervals. It calculates the coordinates on the basis of the analyzed image after sampling. The values of lc, ln, and hc have already been measured. The scan line which is projected from the projector, becomes a real image at point y_i on the base panel in Fig. 4. Point y_i is also measured in section 3.2. However, when the object is placed in front of the base panel, it becomes a real image at point P of the object. It holds the same as the point observed by point y' on the base panel by the web camera. And, the straight line of a, b, and c is linear function, therefore it is easy to estimate an intersection coordinates (x_P, y_P, z_P). We show a schematic diagram of the coordinates calculation in Fig. 4.

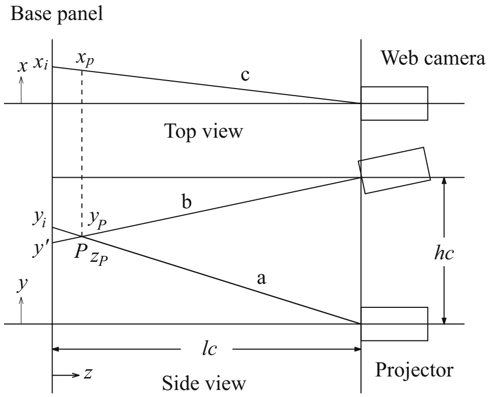


Fig. 4. Computation of coordinate values

$$P = \begin{pmatrix} x_P \\ \\ y_P \\ \\ z_P \end{pmatrix} = \begin{pmatrix} -\dfrac{x_i(y_i - y')}{hc + y_i - y'} + x_i \\ \\ -\dfrac{y_i(y_i - y')}{hc + y_i - y'} + y_i \\ \\ \dfrac{lc(y_i - y')}{hc + y_i - y'} \end{pmatrix} \tag{1}$$

3.5 Conversion of Modeling Data to Polygonal Data

The acquired facial data are converted to polygonal data by a general software application. Representing objects in the form of polygons is a standard modeling technique used in 3DCG as the object can be easily edited. Therefore, we adopt the DXF-3DFACE[7] file format. The simple file structure of DXF simplifies the process of mapping facial data to polygon coordinates. Furthermore, since many software applications support DXF, it is possible to import DXF data from one application to another.

4 Verification, Comparison, and Evaluation of the Acquired Data

4.1 Error Estimation

We calculate the error values. We can see that the accuracy of z_P depends on lc. Therefore, in order to reduce the error in z_P, the value of lc needs to be controlled. We can also see that x_P depends on x_i, which means sampling interval (The number of partitions in vertical direction) of x-axis. Further, y_P (accuracy of projector and web camera) depends on y_i (capture angle).

In our method, equation has to satisfy $y_i - y' > 0$ and $hc > 0$. However, if $y_i - y' > 0$ and $hc > 0$ are not large enough, then it will result in a large error margin.

4.2 Verification of Acquired Data by Geometric Form

It is necessary to measure the dimensions of a known object and verify them with the acquired data to measure the accuracy of the proposed method. In this paper, we verify the accuracy of our proposed method by using a geometric form. A geometric form is a solid model used for sketches etc., whose sizes are known. The accuracy of acquisition data is evaluated by using two geometric forms. The accuracy of this method is verified by comparing the actual dimensional values of the solid model and the measurement data acquired by this method. The actual dimensional values and the measurement data acquired by this method are shown in Fig. 5 and Table 1.

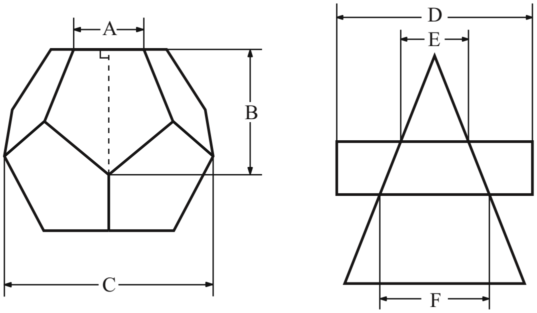


Fig. 5. Dimensions of the solid model

Table 1. Measurement data

	A	B	C
Actual size	70 mm	105 mm	175 mm
Measured value	68 mm	103 mm	171 mm
Difference	-2 mm	-2 mm	-4 mm
	D	E	F
Actual size	185 mm	80 mm	100 mm
Measured value	182 mm	77 mm	98 mm
Difference	-3 mm	-3 mm	-2 mm

4.3 Improved Accuracy of Reconstruction by Using Stereo Matching

We have increased the number of cameras from one to two[2]. Our aim is improved accuracy. We show in Fig. 6 and Fig. 7. We measure the length between the two cameras (lh), and we get values of θ and lt using equation (2). This is the calibration. Using this method we can easily adjust the position, because our method of calibration does not depend on the angle of the cameras. After that, We conduct computation of value from camera A (Fig. 7(a)) and camera B (Fig. 7(b)). We correct the image in camera B (Fig. 7(b)) from equation (3) and (4). (X, Y, Z) are coordinates of a 3D point in the world coordinate space. (u, v) are coordinates of point projection in pixels. (cx, cy) is a principal point (that is usually at the image center). fx and fy are focal lengths expressed in pixel-related units. The joint rotation-translation matrix $[R|t]$ is called a matrix of extrinsic parameters. k_1 and k_2 are radial distortion coefficients, p_1 and p_2 are tangential distortion coefficients. $r^2 = \frac{x}{z}^2 + \frac{y}{z}^2$[3]. We have shown the correct data as image in Fig. 7(c).

$$lt = \frac{lc}{\cos(\arctan \frac{lh}{lc})} = \frac{lh}{\sin(\arctan \frac{lh}{lc})} \tag{2}$$

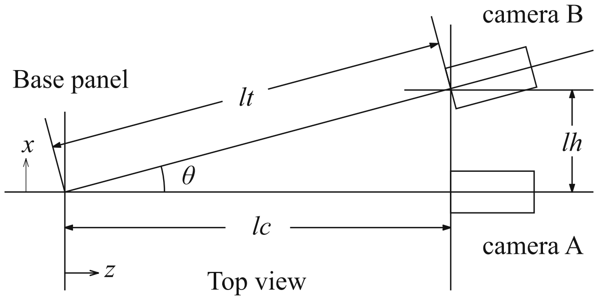


Fig. 6. Two camera setup

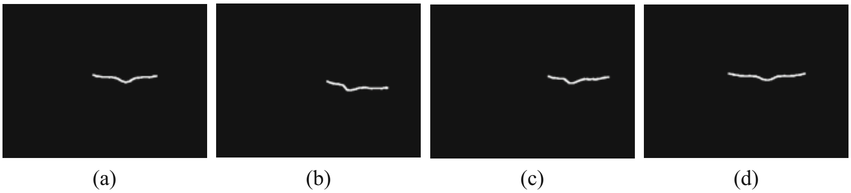


Fig. 7. Image processing to correct two camera

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = R \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} + t \tag{3}$$

$$\begin{pmatrix} u \\ \\ v \end{pmatrix} = \begin{pmatrix} fx[\{\frac{x}{z}(1 + k_1 r^2 + k_2 r^4)\} + \frac{2p_1 xy}{z^2} + p_2\{r^2 + 2(\frac{x}{z})^2\}] + cx \\ fy[\{\frac{y}{z}(1 + k_1 r^2 + k_2 r^4)\} + p_1\{r^2 + 2(\frac{y}{z})^2\} + \frac{2p_2 xy}{z^2}] + cy \end{pmatrix} \tag{4}$$

Then, we synthesize the image in camera A to correct the image (the image in camera B) based on the point of maximum y-scale (that point being the top of nose) to get a more accurate image in Fig. 7(d). We have a satisfactory result as accuracy with two cameras has improved ± 0.5 mm, as compared with one camera. It is higher accuracy than reference [2].

5 Conclusions

In this paper, an economical, simple, and less-time consuming method for generation of photorealistic images is proposed. The image processing method and calculations for this study are described. Furthermore, muscle animation[4] can also be applied if muscular motions of the face are compiled into a database[8]. The simplicity of our method of facial modeling will be applied to the entrance checking information, the interface of the robot and the medical support apparatus etc[11].

References

1. Akimoto, T., Suenaga, Y., Wallace, R., NTT, K.: Automatic creation of 3D facial models. IEEE Computer Graphics and Applications 13(5), 16–22 (1993)
2. An, D., Woodward, A., Delmas, P., Gimelfarb, G., Morris, J.: Comparison of active structure lighting mono and stereo camera systems: Application to 3d face acquisition. In: ENC 2006: Proceedings of the Seventh Mexican International Conference on Computer Science, Washington, DC, USA, pp. 135–141. IEEE Computer Society Press, Los Alamitos (2006)
3. CV Reference Manual, http://opencv.jp/opencv-1.0.0_org/docs/ref/opencvref_cv.htm
4. Bickel, B., Lang, M., Botsch, M., Otaduy, M., Gross, M.: Pose-space animation and transfer of facial details. In: 2008 ACM SIGGRAPH. Eurographics Symposium on Computer Animation (2008)
5. Ghosh, A., Hawkins, T., Peers, P., Frederiksen, S., Debevec, P.: Practical modeling and acquisition of layered facial reflectance. In: SIGGRAPH Asia 2008: ACM SIGGRAPH Asia 2008 papers, pp. 1–10. ACM Press, New York (2008)
6. Lee, Y., Terzopoulos, D., Walters, K.: Realistic modeling for facial animation. In: SIGGRAPH 1995: Proceedings of the 22nd annual conference on Computer graphics and interactive techniques, pp. 55–62. ACM Press, New York (1995)
7. Ochiai, S.: DXF handbook. Ohmsha, Ltd., Tokyo (2003)
8. Platt, S., Badler, N.: Animating facial expressions. ACM SIGGRAPH Computer Graphics 15(3), 245–252 (1981)
9. Smale, S., Zhou, D.: Shannon sampling and function reconstruction from point values. Bulletin-American Mathematical Society 41(3), 279–306 (2004)
10. Terzopoulos, D., Waters, K.: Physically-based facial modeling, analysis, and animation. Journal of visualization and Computer Animation 1(2), 73–80 (1990)
11. Vetter, T., Blanz, V.: A morphable model for the synthesis of 3D faces. In: Proceedings of the ACM SIGGRAPH Conference on Computer Graphics, pp. 187–194 (1999)

Study of Writer Recognition by Japanese Hiragana

Yoshinori Adachi, Masahiro Ozaki, and Yuji Iwahori

Chubu University, 1200 Matsumoto-Cho, Kasugai, Aichi, Japan 487-8501
ozaki@isc.chubu.ac.jp, adachiy@isc.chubu.ac.jp,
iwahori@cs.chubu.ac.jp

Abstract. In a Web study, using user-ID and password to specify learner is common. However, because of disguise problems etc., it is difficult to specify whether the learner on the terminal side is a person in question. Therefore, it is necessary to specify the learner by some other methods. In such methods, there are facial recognition, writer recognition, and other physical recognition methods such as fingerprint, iris, etc. The writer recognition is used for the handwriting analysis well as a concise procedure. In this research, the writer recognition was studied under the characters limited to hiragana. And the similarity differences by character types were examined. Moreover, the differences between similarity values obtained from own dictionary and others' dictionaries were examined. As a result, introducing the stability of the character as one of the identification conditions was found to be necessary. And suitable characters for the writer recognition were obtained. From these results, a new writer recognition method was proposed.

1 Introduction

Japanese is expressed by various characters, i.e. Chinese character, hiragana, katakana, alphabet, etc. Hiragana is a set of characters which can be written from adults to children and is highly used character set in usual occasions. But, it might be understood that the person with frequently use of Chinese character is an educated person. However, there is movement to avoid the use of difficult Chinese characters so as not to cause the misunderstanding that comes from the misuse or miswrite of the Chinese character, and to use hiragana as much as possible.

In **Table 1**, the appearance frequencies of three kinds of character sets in three kinds of books are shown. The appearance of hiragana is more than 60% and the writing frequency of hiragana is much higher in daily life.

We have been studying the writer recognition for a long time[1-9]. In those studies, the Chinese character and the hiragana have been both examined. And the Chinese character showed that the recognition ratio was higher because the Chinese characters have more strokes than the hiragana. Moreover, it was confirmed that there are the hiragana characters not written stably because of the composition of the character, i.e. too simple to write.

However, in the previous study[9], the feature of the hiragana about the writer recognition was studied more precisely, and the possibility to use the hiragana for the writer recognition was shown.

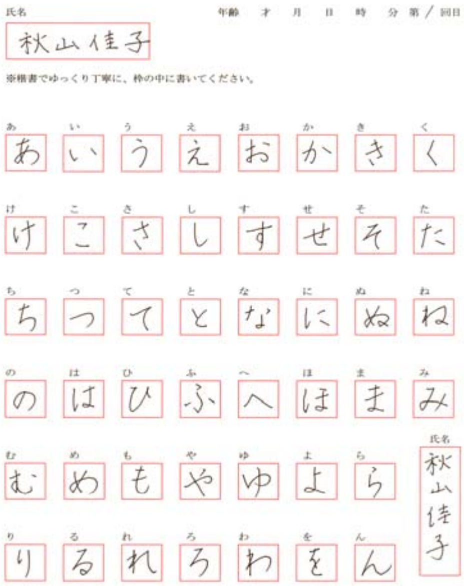
J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 689–696, 2009.

Table 1. Appearance frequency of type of character set in Japanese books

	Chinese	Hiragana	Katakana	Total
Suspense novel	563 (28.8%)	1349 (69.1%)	41 (2.1%)	953
Textbook of mathematics	377 (32.1%)	759 (64.7%)	38 (3.2%)	1174
Science book	358 (34.1%)	665 (63.4%)	26 (0.6%)	1049
Total	1298 (31.1%)	2773 (66.4%)	105 (2.5%)	4176

In this study, to extend the results of the previous study[9], the writer recognition ratio was studied and a new writer recognition method was proposed.

2 Results Obtained in Previous Study



氏名 年齢 才 月 日 時 分 第 / 回目

秋山佳子

※楷書でゆっくり丁寧に、枠の中に書いてください。

あ い う え お か き く

け こ さ し す せ そ た

ち つ て と な に ぬ ね

の は ひ ふ へ ほ ま み

む め も や ゆ よ ら 氏名 秋山佳子

り る れ ろ わ を ん

Fig. 1. Example of filled sheet written by a subject

The special sheet was designed to collect hiragana characters. The size of the frame to write in it was set to be 18mm square. Ten subjects (around 21-years-old) filled out one sheet a day, and they filled ten sheets. **Figure 1** shows the example of the filled sheet.

After the sheet image was input to a computer through a scanner (280dpi), each character was cut out one by one based on the each frame automatically. Therefore, the number of character samples became 4600 (10 subjects x 46 types x 10 times) in total.

The writer's feature was extracted by using the new local arc method. The chord length was adopted 13 dots which gave the highest accuracy in the last study [8]. The angle of the chord had been changed from 0° to 180° at every 15°. A curvature of the stroke of the character was obtained within the range from -5 to 5, and an appearance frequency of curvature was obtained as a feature vector of the character in 12x11 dimensions (12 angles x 11 curvatures).

The dictionary of each subject and each type of character was made as follows. First of all, the feature vectors of each type of character in 5 sheets were resolved to the eigenvalues and the eigenvectors by the principal component analysis. The eigenvectors were added so that the accumulation contribution rate of the eigenvalues might become 90% or more. The weighted average of the feature vectors of 5 characters was calculated from the elements of the added eigenvector. In this study, the accumulation contribution rate of 90% or more was achieved by two eigenvalues. The similarity value was calculated by the commonly used cosine value with the dictionary, i.e. inner product. The mean value of 5 similarity values was assumed to be a similarity value of the character, and it was obtained for each type of character and each subject. Moreover, similarity values were also obtained from other's dictionary, and the mean values obtained from own dictionary and other's dictionaries were compared each other. It was examined whether there was a difference in the average similarity values by t-test.

In **Figure 2**, the characters are arranged in the order of similarity values, and also in **Figure 3**, the characters are arranged in the order of t-values. The orders are completely different.

In the previous study[9], large t-value characters were recommended as appropriate characters for identification, i.e. characters "そよくわまやなねめを" were recommended.

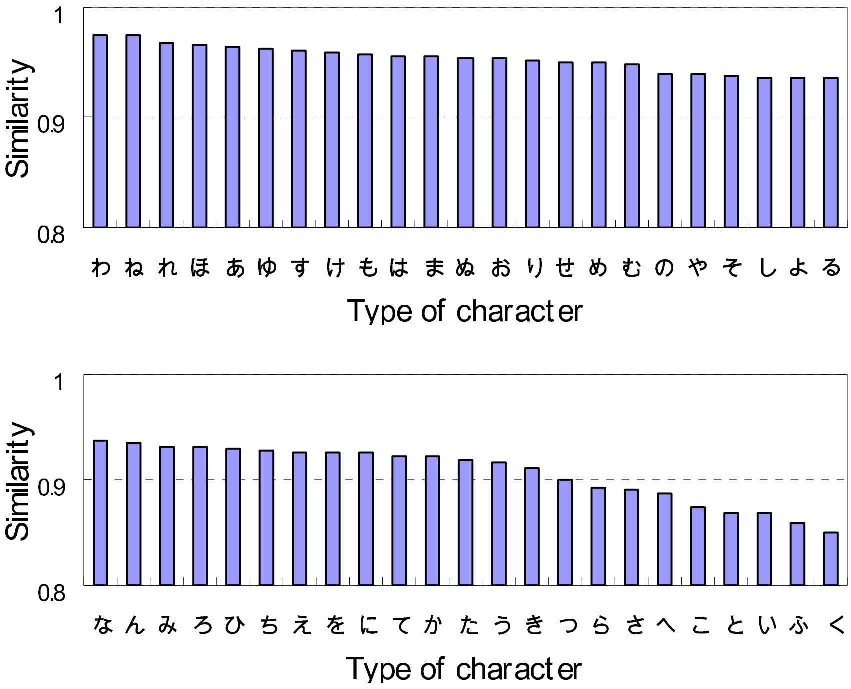


Fig. 2. Similarity values vs. character type

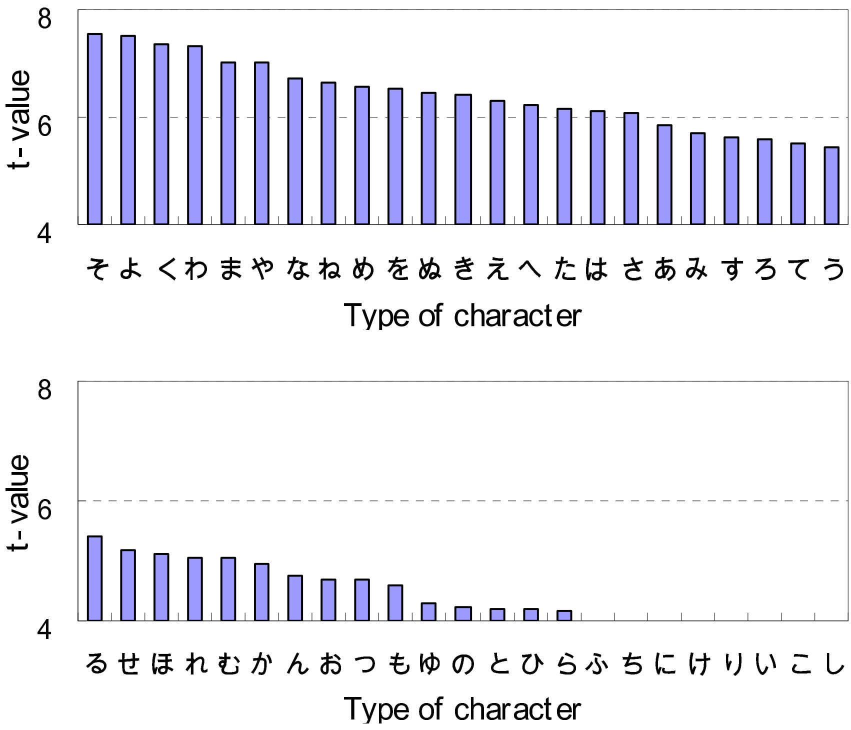


Fig. 3. t-values vs. character type

3 Experimental Results

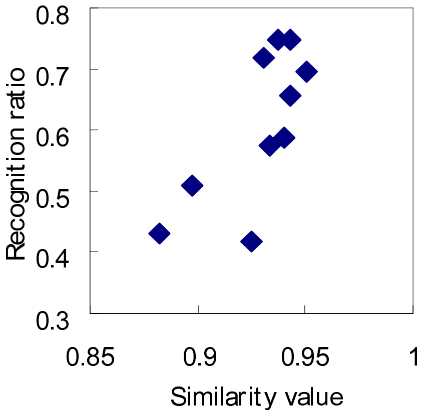


Fig. 4. Relation between similarity value vs. recognition ratio of each subject

First of all, the relation between the average similarity value of the each type of characters and writer recognition ratio was examined. **Figure 4** shows the relation between the average similarity value and the recognition ratio of each subject. And **Figure 5** shows that of each type of character. In Figure 4, the correlation coefficient is r=0.721 and it shows that the recognition ratio is fairly related to the similarity

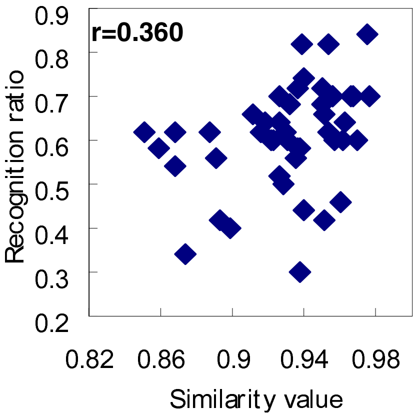


Fig. 5. Relation between similarity value vs. recognition ratio of each character

value. It is thought to come from the stability of writing characters.

On the other hand, Figure 5 shows that the recognition ratio of large similarity value character is not necessarily large, i.e. r=0.360. In a word, it is indicated that large similarity value characters are not necessarily appropriate for the writer recognition. However, the similarity value shows the stability of the character, and should choose characters those similarity values are as large as possible.

Next, the relation between t-value and recognition ratio of each character is shown in **Figure 6**. This figure shows fairly strong relation between t-value and recognition ratio, i.e. r=0.745. In a word, large t-value characters can be chosen as appropriate characters for the writer recognition. Therefore, in this study, large t-value characters those similarity values are larger than the average similarity value are selected as appropriate characters for the writer recognition.

Therefore, the characters "そよわまやなねめ" are recommended from Figures 2 and 3. However, those recognition ratios are not so large as listed in **Table 2**.

Therefore, in this study, we proposed a new evaluation function expressed by the following equation consisted of 4 similarity values.

$$\eta = 1 - \left[(1-\eta_1)(1-\eta_2)(1-\eta_3)(1-\eta_4)\right]^{\frac{1}{4}} \tag{1}$$

where η is the evaluation value, and η_i (i=1,2,3,4) is one of the similarity values of four types of characters, i.e. "そよわま", respectively.

As a result, the recognition ratio became 100%.

Table 2. Recognition ratio of large t-value characters

Type of character	Recognition ratio	Type of character	Recognition ratio
そ	0.82	や	0.74
よ	0.58	な	0.58
わ	0.7	ね	0.84
ま	0.7	め	0.72

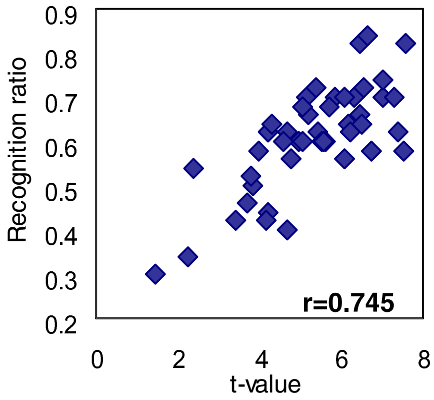


Fig. 6. Relation between t-value vs. recognition ratio of each character

To show the importance of t-value, the recognition ratios are calculated from the similar evaluation function as Eq.(1) with “しこいり”, and are listed in **Table 3**. The results indicate the importance of t-value.

4 Proposed Writer Recognition Process

In this study, we found the suitable characters and the evaluation function for the writer recognition. Then, we propose a new writer recognition process shown in **Figure 7**.

Table 3. Comparison of recognition ratio obtained from large t-value characters and small t-value characters with Eq.(1)

	Large t-value characters	Small t-value characters
Recognition ratio	100%	54%

5 Conclusion

From the above-mentioned results, we obtained the followings:

(1) The characters which have large similarity values are not necessarily suitable for the writer recognition.

(2) Even though the similarity values are small, there are characters which express writer’s features well and have large t-values.

(3) We recommended characters “そよわま” for the writer recognition together with the evaluation function :

$$\eta = 1-\left[(1-\eta_1)(1-\eta_2)(1-\eta_3)(1-\eta_4)\right]^{\frac{1}{4}}$$

In the present study, number of subject was only 10, and then the results might not be quite accurate. In the future, it will be necessary to increase the number of writers, and to examine the influence of the type of character on the writer recognition further.

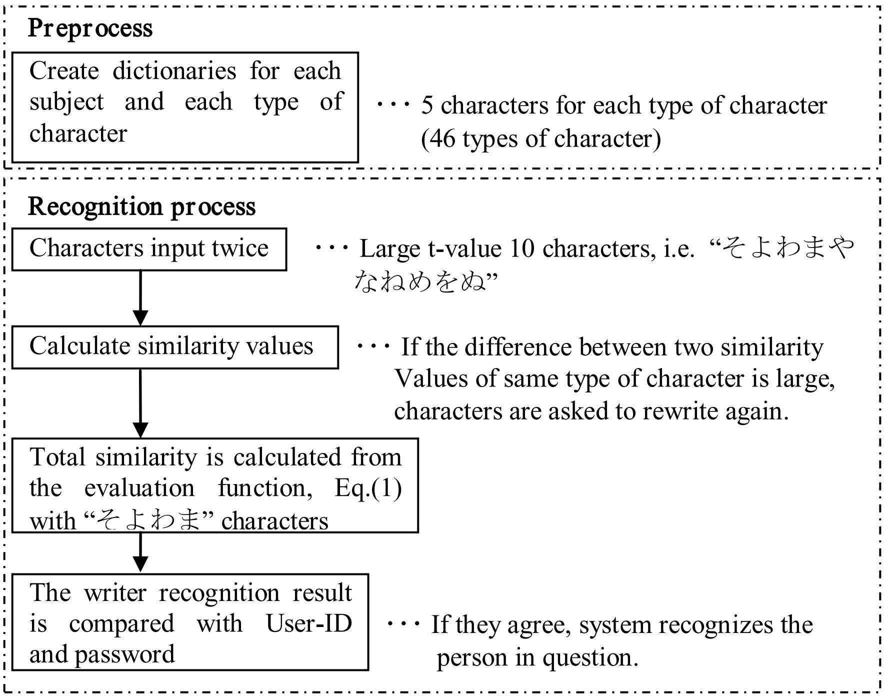

Preprocess
Create dictionaries for each subject and each type of character
··· 5 characters for each type of character (46 types of character)
Recognition process
Characters input twice
··· Large t-value 10 characters, i.e. "そよわまやなねめをぬ"
Calculate similarity values
··· If the difference between two similarity Values of same type of character is large, characters are asked to rewrite again.
Total similarity is calculated from the evaluation function, Eq.(1) with "そよわま" characters
The writer recognition result is compared with User-ID and password
··· If they agree, system recognizes the person in question.

Fig. 7. Flow of identification process

Acknowledgment

This research is supported by JSPS Grant-in-Aid for Scientific Research, Scientific Research (C) (19500830) and Chubu University Grant.

References

[1] Ozaki, M., Adachi, Y., Ishii, N., Koyazu, T.: Fuzzy CAI System to Improve Hand Writing Skills by Using Sensuous. Trans. of IEICE J79-D(9), 1554–1561 (1996)

[2] Ozaki, M., Adachi, Y., Ishii, N.: Writer Recognition by means of Fuzzy Similarity Evaluation Function. In: Proc. KES 2000, pp. 287–291 (2000)

[3] Ozaki, M., Adachi, Y., Ishii, N.: Study of Accuracy Dependence of Writer Recognition on Number of Character. In: Proc. KES 2000, pp. 292–296 (2000)

[4] Ozaki, M., Adachi, Y., Ishii, N., Yoshimura, M.: Writer Recognition by means of Fuzzy Membership Function and Local Arcs. In: Proc. KES 2001, pp. 414–418 (2001)

[5] Ozaki, M., Adachi, Y., Ishii, N.: Development of Hybrid Type Writer Recognition System. In: Proc. KES 2002, pp. 765–769 (2002)

[6] Adachi, Y., Liu, M., Ozaki, M.: A New Similarity Evaluation Function for Writer Recognition of Chinese Character. In: Negoita, M.G., Howlett, R.J., Jain, L.C. (eds.) KES 2004. LNCS, vol. 3214, pp. 71–76. Springer, Heidelberg (2004)

[7] Ozaki, M., Adachi, Y., Ishii, N.: Writer Recognition by Using New Searching Algorithm in New Local Arc Method. In: Khosla, R., Howlett, R.J., Jain, L.C. (eds.) KES 2005. LNCS, vol. 3681, pp. 775–780. Springer, Heidelberg (2005)

[8] Adachi, Y., Ozaki, M., Iwahori, Y., Ishii, N.: Influence of presence of frame on writer recognition. In: Apolloni, B., Howlett, R.J., Jain, L. (eds.) KES 2007, Part II. LNCS, vol. 4693, pp. 1045–1050. Springer, Heidelberg (2007)

[9] Adachi, Y., Ozaki, M., Iwahori, Y.: Influence of character type of japanese hiragana on writer recognition. In: Lovrek, I., Howlett, R.J., Jain, L.C. (eds.) KES 2008, Part II. LNCS, vol. 5178, pp. 934–941. Springer, Heidelberg (2008)

Speed Flexibility Biomedical Vision Model Using Analog Electronic Circuits and VLSI Layout Design

Masashi Kawaguchi[1], Shoji Suzuki[1], Takashi Jimbo[2], and Naohiro Ishii[3]

[1] Department of Electrical & Electronic Engineering, Suzuka National College of Technology, Shiroko, Suzuka Mie 510-0294, Japan
{masashi,shoji}@elec.suzuka-ct.ac.jp
[2] Department of Environmental Technology and Urban Planning Graduate School of Engineering, Nagoya Institute of Technology,
Gokiso-cho, Showa-ku, Nagoya, 466-8555 Japan
jimbo.takashi@nitech.ac.jp
[3] Department of Information Science, Aichi Institute of Technology,
Yachigusa, Yagusa-cho, Toyota, 470-0392 Japan
ishii@aitech.ac.jp

Abstract. We propose here an artificial vision model for the speed flexibility motion detection which uses analog electronic circuits and design the analog VLSI layout. In the previous model, the range of speed is quite narrow. However, we use the variable resistant parts inside the circuits. This model has speed flexibility property, and it is comprised of four layers. The model was shown to be capable of detecting a movement object. The number of elements in the model is reduced in its realization using the integrated devices. Therefore, the proposed model is robust with respect to fault tolerance. Moreover, the connection of this model is between adjacent elements, making hardware implementation easy.

Keywords: Neural Network, Motion Detection, Analog Circuits, Biomedical Vision System.

1 Introduction

A neuro chip and an artificial retina chip are developed to comprise the neural network model and simulate the biomedical vision system. At present, a basic image processing, such as edge detection and reverse display of an image has been developed [1][2]. The retina consists of the inside retina and outside retina. The inside retina sends the nerve impulses to the brain, whereas the outside retina receives optical input from the visual cell. As a result, the outside retina emphasizes spatial changes in optical strength. Recently, the network among the amacrine cell, the bipolar cell and the ganglion cell has been clarified theoretically, which has led to active research concerning the neuro-device, which models the structure and function of the retina. Easy image processing, reversing, edge detection, and feature detection, have been achieved by technologies such as the neuro chip and the analog VLSI circuit.

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 697–704, 2009.

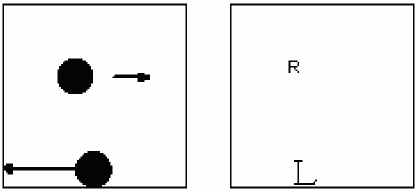


Fig. 1. Example of Advanced Image Processing

Some motion detection models are proposed in the recent researches. Figure 1 shows the example of advanced image processing. It is direction sensitive motion detection behavior. When the object moves from left to right slowly the model outputs a small "right" signal, and when the object moves from right to left quickly the model outputs a big "left" signal.

2 Advanced Image Processing

Lu et al. describes the application of an analog VLSI vision sensor to active binocular tracking. The sensor outputs are used to control the vergence angles of the two cameras and the tilt angle of the head so that the center pixels of the sensor arrays image the same point in the environment[3]. Another model presents the implementation of a visual motion detection algorithm on an analog network. The algorithm in the model is based on Markov random field (MRF) modeling. Robust motion detection is achieved by using a spatiotemporal neighborhood for modeling pixel interactions. Not only are the moving edges detected, but also the inner part of moving regions [4]. The other model is an analog MOS circuit inspired by an inner retina. The analog circuit produces signals of motion of edges which are output in an outer retinal neural network. Edge signals are formed into half-wave rectified impulses in two types of amacrine cells, and fed back to the wide field amacrine cell in order to modulate width of impulses [5]. However, these models can detect the movement direction only or speed only. In the present study, we propose a motion detection model in which the speed is detected by differentiation circuits.

3 One Dimensional Motion Detection Model

We first developed a one-dimensional model, the structure of which is shown in Fig. 2. The surface layer is composed of the connections of capacitors. In the inner layer, the movement direction is detected by difference circuits. When the object moves from left to right, a positive output signal is generated, and when the object moves from right to left, a negative output signal is generated. We show this model is able to detect the speed and direction of a movement object by the simple circuits. Despite the large object size, this model can detect the motion.

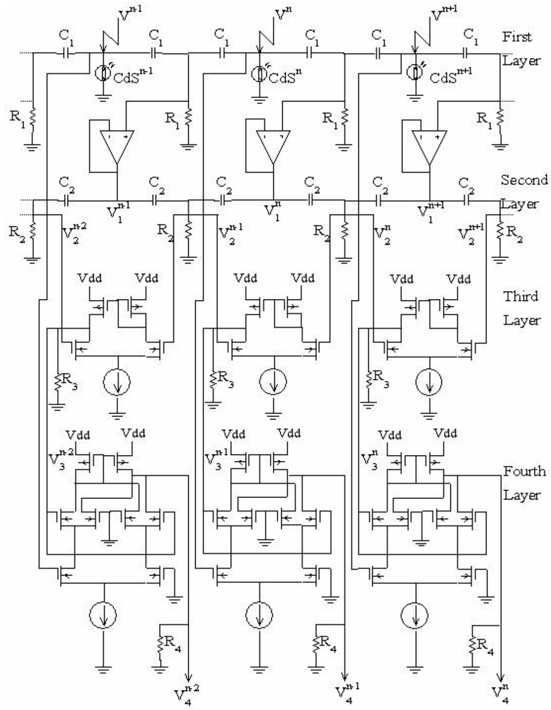


Fig. 2. One-Dimensional Four-Layered Direction Model for Selective Motion Detection

3.1 First Layer Differentiation Circuits (First Layer)

The current is given by equation (1), where the input voltage is denoted by V^n and the capacitance is denoted by C_1. The current into a capacitor is the derivative with respect to time of the voltage across the capacitor, multiplied by the capacitance.

$$I = C_1 \frac{dV^n}{dt} \tag{1}$$

$$V_1^n = IR_1 = C_1 R_1 \frac{dV^n}{dt} \tag{2}$$

The output voltage V_1^{n} is given by equation (2). Equation (2) is multiplied by the resistance R_1, calculating the voltage potential. Buffer circuits are realized by operational amplifiers between the first layer and the second layer. In the first layer, there are also the CdS Photoconductive Cells. Using CdS cells, this model is not affected by object luminance. When the object is high luminance, the resistances of CdS cells are low. Some currents flows to ground through the CdS. Therefore, despite the high luminance, the input Voltage V_1^{n} is not affected.

3.2 Second Layer Differentiation Circuits (Second Layer)

The second layer is also composed of differentiation circuits; however, the CR coefficient is small compared that of the first layer differentiation circuits. The output of first layer, V_1^n, is differentiated again, and the output of the second layer is assumed to be V_2^n, calculating the voltage potential.

$$I = C_2 \frac{dV_1^n}{dt} \tag{3}$$

$$V_2^n = IR_2 = C_2 R_2 \frac{dV_1^n}{dt} \tag{4}$$

3.3 Difference Circuits (Third Layer)

The third layer consists of difference circuits realized by MOSFET. The bottom I_b is a current source. The manner in which I_b is divided between Q_1 and Q_2 is a sensitive function of the difference between V_2^{n+1} and V_2^n, and is the essence of the operation of the stage. We assume the MOSFET device is in the sub-threshold region and the *I-V* characteristics follows the exponential characteristics, then the drain current I_D in the sub-threshold region is exponential in the gate voltage V_g and source voltage V_s. *V* is electric potential of current source I_b. I_0 and κ are coefficients.

The circuit consists of a differential pair and a single current mirror, like the one shown in Figure 2, which is used to subtract the drain currents I_1 and I_2. The current I_1 drawn out of Q_3 is reflected as an equal current out of Q_4; the output current I_{out} is thus equal to I_1- I_2, and is therefore given by Equation (5).
The output voltage of this circuit is as follows.

$$\begin{aligned} I_1 - I_2 &= I_b \frac{e^{\kappa\left(V_2^{n+1}-V_2^n\right)/2} - e^{-\kappa\left(V_2^{n+1}-V_2^n\right)/2}}{e^{\kappa\left(V_2^{n+1}-V_2^n\right)/2} + e^{-\kappa\left(V_2^{n+1}-V_2^n\right)/2}} \\ &= I_b \tanh \frac{\kappa(V_2^{n+1} - V_2^n)}{2} \end{aligned} \tag{5}$$

$$V_3^n = (I_1 - I_2)R_3 = I_b R_3 \tanh \frac{\kappa(V_2^{n+1} - V_2^n)}{2} \tag{6}$$

3.4 Gilbert Multiple Circuits (Fourth Layer)

The fourth layer is comprised of Gilbert multiple circuits. We assume the MOSFET device is in the sub-threshold region, the *I-V* characteristics follows the exponential characteristics, then the drain current I_D in the sub-threshold region is exponential in the gate voltage V_g and source voltage V_s. The results for the two drain currents of the differential pair were derived. In Figure 2, V_1 and V_2 are connected to ground respectively. In this circuit, the voltage V_1 and V_2 are 0. The fourth layer produces the third layer output V_3^n and the input signal V^{n+1}. This circuit detects the pure output of movement. I_b is the current source, and κ is a coefficient [1].

$$I_4^n = I_b \tanh\frac{\kappa V_3^n}{2} \tanh\frac{\kappa V^{n+1}}{2} \tag{7}$$

$$V_4^n = I_4^n R_4 = I_b R_4 \tanh\frac{\kappa V_3^n}{2} \tanh\frac{\kappa V^{n+1}}{2} \tag{8}$$

I_4^n is the output current of the fourth layer, R_4 is the earth resistance, and V_4^n is the final output. I_4^n corresponds to I_{out}. Using multiple circuits, this model can detect the pure output of movement. We set the parameter of circuits as follows. In the first layer, C_1=0.1μF, R_1=1kΩ. We used the μA741 as a buffer circuits. In the second layer, C_2=0.1μF, R_2=100kΩ. At the difference circuits, we used the VP1310 and VN1310 as MOSFET [6]. The connection of this model is between adjacent elements, making hardware implementation easy. We measured the shape of the output waves produced by the input movement signal using an electronic circuit simulator (SPICE).

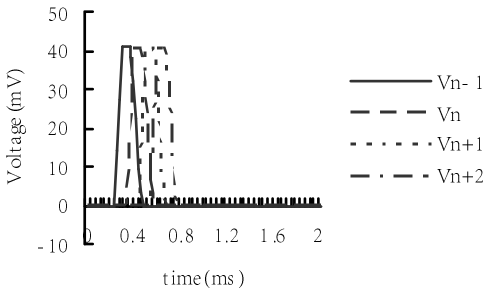


Fig. 3. Output of the fourth layer(after multiple circuit processing)

Figure 3 shows the final output of the forth layer when the object moves from left to right, which indicates that this circuit detects the pure output of the movement. We have other experiments. When the object moves at half speed, this model outputs a lower signal. When the object moves from right to left, this model outputs a negative signal. This model can detect the speed and direction of a movement object in one dimension [7][8].

3.5 Improvement in Motion Detection Model

This model has some problems. One is the detecting range of speed is limited. When the moving object speed is about 10 segments per seconds, this model can detect the movement. That is, this model was designed as the speed flexibility system. In the biomedical devices, used for vision and brain, the flexibility system is working. We used coupled LED and CdS photo-resisters for speed flexibility model. We show this parts in Figure 4. When the input voltage is low, the luminance of LED is low.

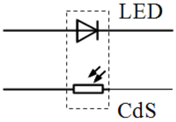
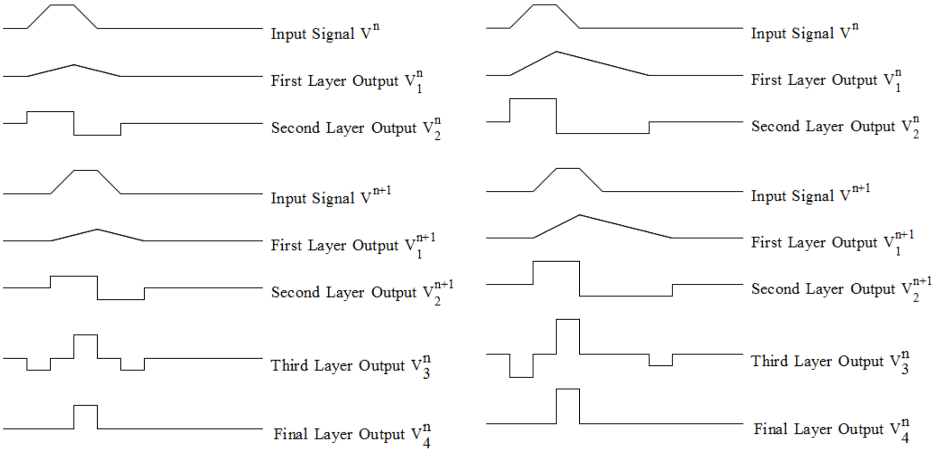


Fig. 4. LED-CdS photo-resister


Input Signal V^n
First Layer Output V_1^n
Second Layer Output V_2^n
Input Signal V^{n+1}
First Layer Output V_1^{n+1}
Second Layer Output V_2^{n+1}
Third Layer Output V_3^n
Final Layer Output V_4^n
Input Signal V^n
First Layer Output V_1^n
Second Layer Output V_2^n
Input Signal V^{n+1}
First Layer Output V_1^{n+1}
Second Layer Output V_2^{n+1}
Third Layer Output V_3^n
Final Layer Output V_4^n

previous model[7][8] proposed model

Fig. 5. Diagram of each layer and each terminal

However, when the input voltage is high, the luminance of LED is high. On the other hand, we describe CdS photo- when the luminance is high, the resistance of CdS is low.

By combination of LED-CdS photo-resisters, when the input voltage is low, the resistance of "LED-CdS photo-resisters" is high. However, when the input voltage is high, the resistance of "LED-CdS photo-resisters" is low.We connected this LED-CdS photo-resister to output layer and first layer. In case the object is stopping or moving very slowly, the resistance of photo-resisters becames maximum and CR coefficient of first layer becomes very large. This previous model can detect the slow movement only.

On the other hand, object moving quickly, the final output becomes large value. Thus, photo resisters value becomes low. In this situation, CR coefficient value is small and capacitor is charged quickly. After passing moving object, final output becomes small value and CR coefficient value becomes large. It is keeping the output value of the first layer. When the object moves very quickly, object is missed in the previous model. Because it cannot keep the output value of the first layer until appearing signal from the neighbor first layer. However, in the proposed model, object will not be missed in case of the quick movement.

Figure 5 shows the input and output signal of each layer. The output of the first layer indicates that input signal is differentiated by a large CR coefficient. The output of the second layer shows that the first layer output signal is differentiated by a small CR coefficient. V^{n+1} indicates the neighbor terminal of V^n. Therefore, the input signal of V^{n+1} is delayed compared to V^n. The difference between the second layer output V^{n+1} and V^n is calculated. The third layer output shows the peak positive signal. Finally, the fourth layer produces the third layer output V_3^n and the input signal V^{n+1}. This circuit detects the pure output of movement. The proposed model can detect the high level output compare with the previous model shown in Figure 5. However, we proposed only one-dimensional model. The two-dimensional model using LED-CdS photo-resisters is future work.

This behavior is also realized using double integrated A/D translate circuit. However, the circuit structure of the proposed model is very simple and easy to make hardware.

Moreover, in the future, RRAM (Resistance RAM), SMRE (Semiconductor Magnetoresistive Elements) or memristor (memory resistor) parts will be developed. These parts are very small in size and simple structure compared to the LED-CdS photo-resister.

3.6 Designs for Circuit Board

Next, we designed the circuit board using CAD system developed by MITS Corporation. This data is for making the circuit board using manufacturing system. In this paper, we show that it is realized that this model is by the real circuit, not by simulation.

3.7 Layout for Motion Detection Circuits

The proposed model is processed by the analog electronic circuits. We designed the simulated circuit to the chip layout using Orcad Layout Tool. We show that it is possible to realize the hardware implementation on the integrated circuits. In the biomedical brain, information is also processed in an analog manner. In the future, movement information will be collected into an analog electronic brain model. This would allow the hardware system of the biomedical brain model to be realized. The proposed moving detection model has possible application as a sensor and can composed part of the receptor. The proposed model will enable the clarification of the mechanism of the biomedical brain.

4 Conclusion

We designed the motion detection analogue electric circuit using a biomedical vision system. We first designed the one-dimensional model and experimented. Using the one-dimension model, the movement information was detected. The input terminal and the output terminal were arranged in an alternating manner. As a result, a simple circuit and an equivalent output result were obtained. The realization of an integration device will enable the number of elements to be reduced. The proposed model is robust with respect to fault tolerance, in case the extra output has been generated, if the scale of the model is enhanced comprehensively, this will not present a significant problem. Moreover, the connection of this model is between adjacent elements, making hardware implementation easy. Finally, we designed the layout of analog VLSI model. We show that its model is possible to realize the hardware implementation[7].

Acknowledgments. This research was financially supported by Grant-in-Aid from the Japanese Ministry of Education, Science, Sports and Culture.

References

1. Mead, C.: Analog VLSI and Neural Systems. Addison Wesley Publishing Company, Inc, Reading (1989)
2. Chong, C.P., Salama, C.A.T., Smith, K.C.: Image-Motion Detection Using Analog VLSI. IEEE Journal of Solid-State Circuits 27(1), 93–96 (1992)
3. Lu, Z., Shi, B.E.: Subpixel Resolution Binocular Visual Tracking Using Analog VLSI Vision Sensors. IEEE Transactions on Circuits and Systems-II: Analog and Digital Signal Processing 47(12), 1468–1475 (2000)
4. Luthon, F., Dragomirescu, D.: A Cellular Analog Network for MRF-Based Video Motion Detection. IEEE Transactions on Circuits and Systems-I: Fundamental Theory and Applications 46(2), 281–293 (1999)
5. Yamada, H., Miyashita, T., Ohtani, M., Yonezu, H.: An Analog MOS Circuit Inspired by an Inner Retina for Producing Signals of Moving Edges. Technical Report of IEICE, NC99-112, pp.149–155 (2000)
6. Kawaguchi, M., Jimbo, T., Umeno, M.: Motion Detecting Artificial Retina Model by Two-Dimensional Multi-Layered Analog Electronic Circuits. IEICE Transactions E86-A-2, 387–395 (2003)
7. Kawaguchi, M., Jimbo, T., Umeno, M.: Analog VLSI Layout Design of Advanced Image Processing For Artificial Vision Model. In: IEEE International Symposium on Industrial Electronics, ISIE 2005 Proceeding, vol. 3, pp. 1239–1244 (2005)
8. Kawaguchi, M., Jimbo, T., Umeno, M.: Analog VLSI Layout Design and the Circuit Board Manufacturing of Advanced Image Processing for Artificial Vision Model. In: Lovrek, I., Howlett, R.J., Jain, L.C. (eds.) KES 2008, Part II. LNCS, vol. 5178, pp. 895–902. Springer, Heidelberg (2008)

Self-calibration and Image Rendering Using RBF Neural Network

Yi Ding[1], Yuji Iwahori[2], Tsuyoshi Nakamura[1], Robert J. Woodham[3], Lifeng He[4], and Hidenori Itoh[1]

[1] Dept. of Computer Science and Eng.,
Nagoya Institute of Technology
Gokiso-cho, Showa-ku Nagoya 466-8555, Japan
ding@juno.ics.nitech.ac.jp, {tnaka, itoh}@nitech.ac.jp
http://www.nitech.ac.jp/

[2] Dept. of Computer Science, Chubu University
Matsumoto-cho 1200, Kasugai 487-8501, Japan
iwahori@cs.chubu.ac.jp
http://www.cvl.cs.chubu.ac.jp

[3] Dept. of Computer Science, University of British Columbia
Vancouver, B.C. Canada, V6T 1Z4
woodham@cs.ubc.ca
http://www.cs.ubc.ca/~woodham/

[4] Faculty of Information Science and Technology,
Aichi Prefectural University, Aichi 480-1198, Japan
helifeng@ist.aichi-pu.ac.jp
http://www.aichi-pu.ac.jp

Abstract. This paper describes a new approach for self-calibration and color image rendering using radial basis function (RBF) neural network. Most empirical approaches make use of a calibration object. Here, we require no calibration object to both shape recovery and color image rendering. The neural network training data are obtained through the rotations of a target object. The approach can generate realistic virtual images without any calibration object which has the same reflectance properties as the target object. The proposed approach uses a neural network to obtain both surface orientation and albedo, and applies another neural network to generate virtual images for any viewpoint and any direction of light source. Experiments with real data are demonstrated.

Keywords: Neural Network Based Rendering, Photometric Stereo, Self-Calibration, Albedo, Shape Recovery.

1 Introduction

Model based rendering purposes generating realistic images from the 3-D modeling of the real object. In general, 3-D modeling deals with both photometric and geometric properties such as shape, viewpoint, lighting, and albedo. Rendering is originally based on the technology of 3-D computer graphics and it has been used in graphics architecture, video games and recently in the area of computer vision and mixed reality.

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 705–712, 2009.

In the previous approaches for shape recovery, Woodham [1] proposed a method to recover the surface orientation from shading images using photometric stereo. Further, an empirical approach to photometric stereo was proposed in [2]. Empirical photometric stereo uses a calibration sphere with the same reflectance properties as the target object.

Iwahori *et al.* [3] developed neural network implementations of photometric stereo. Neural network based photometric stereo learns the surface reflectance property by learning the mapping of triples of image irradiance to the corresponding surface orientation, with surface reflectance factor [4], using a calibration sphere.

The approach [5] uses the reflectance factor and surface normal vector for the color image rendering using a calibration sphere. The 3D shape model and color reflectance factor are obtained by neural network, then the approach is applied for neural network based image rendering to generate virtual images for arbitrary viewpoint and direction of light source.

In this paper, we propose a new approach to improve neural network based rendering without any calibration sphere. Instead, the rotation of the target object itself generates the learning data for neural network via self-calibration. Using the dichromatic reflection model, the image irradiances of specular reflection and diffuse reflection are separated from the observed data. Four images of target object under different four light sources are used to recover the shape and to generate virtual images, including recovering color reflectance factor.

2 Background

2.1 Principle of Photometric Stereo

In [4], neural network based photometric stereo is used to determine both surface orientation and surface albedo. In four light source neural network based empirical photometric stereo using a calibration sphere which has the same reflectance properties as the target object, the following constraint equation holds.

$$\begin{cases} E_1(x,y) = R_1(\boldsymbol{n},\rho) \\ E_2(x,y) = R_2(\boldsymbol{n},\rho) \\ E_3(x,y) = R_3(\boldsymbol{n},\rho) \\ E_4(x,y) = R_4(\boldsymbol{n},\rho) \end{cases} \tag{1}$$

where (R_1, R_2, R_3, R_4) is the reflectance map, $\boldsymbol{n}$ is the surface normal vector and ρ represents the reflectance factor (albedo). In the previous approaches [4], neural network learns the mapping of (E_1, E_2, E_3, E_4) to $(\boldsymbol{n}, \rho)$ for a calibration object.

2.2 Dichromatic Reflection Model

The Dichromatic Reflection Model describes that image irradiance E consists of two components, one is a diffuse component R_d and the other is a specular component R_m. The parameters d (diffuse component) and m (specular component) represent the mixing ratios of the dichromatic reflection model. The mixing ratio depends on the surface normal vector $\boldsymbol{n}$, the light source direction $\boldsymbol{s}$ and the viewing direction $\boldsymbol{v}$ at each point on the object. E_w represents the intensity of light source.

$$E = dE_w\rho + mE_w = R_d\rho + R_m \tag{2}$$

3 Self-calibration and Neural Network Learning

3.1 Self-calibration with Rotation

The self-calibration is to obtain the training data for a neural network from the target object itself. The target object is rotated from 0 to 359 degree. The object images are obtained under each of four light sources. At the rotation angle 0 degree, feature points on the occluding boundaries are extracted. These feature points are tracked during the rotation and the data obtained are used for neural network training.

Gaussian sphere is defined as a virtual sphere with its radius $R = 1$. Each point on the Gaussian sphere is projected onto the tangent space defined (f, g) space with the stereographic projection. Although $(p, q) = (\frac{\delta Z}{\delta X}, \frac{\delta Z}{\delta Y})$ becomes infinity on the occluding boundaries, (f, g) values takes within a circle of radius 2 for all points [6].

During rotation, the feature points on the occluding boundaries are selected and tracked with every 1 degree. Suppose a feature point on the occluding boundaries, tracking points from this point during rotation is located inside the circle of radius 2 in (f, g) plane. The radius of the circle, r, represents the horizontal distance at each point from the boundary to the rotation axis, where $r = Rcos\theta = cos\theta$.

For the feature point $a(x_a, y_a)$, two pixels lower point $b(x_b, y_b)$ and two pixels upper point $c(x_c, y_c)$ of the point $a(x_a, y_a)$ are used and surface orientation (f, g) of each feature point is obtained by the geometrical approximation of small triangle on the occluding boundaries.

Using (f, g) representation, the current feature point is determined from the current rotation angle α, R and r as

$$(f, g) = (\frac{2Rrcos\alpha}{R + rsin\alpha}, \pm\frac{\sqrt{(f^2 + 4R^2)(R^2 - r^2)}}{\sqrt{(R + rsin\alpha)^2 + (rcos\alpha)^2}}) \tag{3}$$

The corrspongind gradient parameters (p, q) is given by

$$p = \frac{4f}{4 - f^2 - g^2}, \quad q = \frac{4g}{4 - f^2 - g^2} \tag{4}$$

for points where $(4 - f^2 - g^2) \neq 0$. The surface normal, $(\boldsymbol{n}_x, \boldsymbol{n}_y, \boldsymbol{n}_z)$, is computed from (p, q) as

$$(\boldsymbol{n}_x, \boldsymbol{n}_y, \boldsymbol{n}_z) = \frac{(-p, -q, 1)}{\sqrt{p^2 + q^2 + 1}} \tag{5}$$

3.2 NN Implementation for Shape Recovery

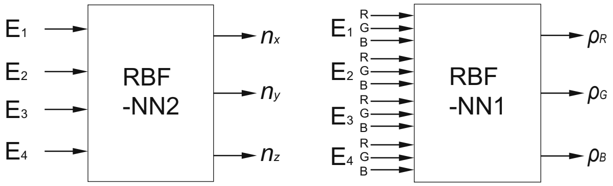
Based on selected feature points which construct a virtual sphere, the mapping of (E_1, E_2, E_3, E_4) to (n_x, n_y, n_z) is used as the training data under four light sources.

In previous neural network based photometric stereo, a calibration object should be composed of exactly the same material as the target object. The number of training data increases when we use color in NN, because a color image has more information than a grayscale image. Further, random color reflectance factor, $\boldsymbol{\rho}^{'} = (\rho^{'}_R, \rho^{'}_G, \rho^{'}_B)$, is synthesized for each point on the training data, $(E^{'}_R, E^{'}_G, E^{'}_B) = (\rho^{'}_R E_R, \rho^{'}_G E_G, \rho^{'}_B E_B)$ to estimate the reflectance factor first. The diffuse component and the specular component are separated in Equation (2), Here, only the diffuse component is synthesized and added for each point on the training data by $\boldsymbol{\rho}^{'}$.

To learn the NN efficiently, the highest value is chosen and used as the monochrome image intensity E from the color values $(E_R,\ E_G,\ E_B)$ included in the target object.

$$E = \max(E_R, E_G, E_B) \tag{6}$$

In the learning during self-calibration, the training data obtained from the feature points is chosen using Equation (6) and (E_1, E_2, E_3, E_4) are given as the input to NN. The corresponding (n_x, n_y, n_z) are given as the output of NN. After learning of the NN, the input data (E_1, E_2, E_3, E_4) of a test object are given, then, the corresponding (n_x, n_y, n_z) of the test object as the output data is obtained from NN in generalization for the test object. The structure of this RBF-NN (Radial Basis Function Neural Network) is shown in Fig.1-(a).



(a) NN for surface normal estimation

(b) NN for color reflectance factor estimation

Fig. 1. RBF-NN for Self-Calibration

3.3 Color Reflectance Factor

According to Eq.(2), the reflectance factor of any feature point is calculated using the following equation.

$$(E_R - E_{Rm})/E_{Rd} = \rho_R \tag{7}$$

$$(E_G - E_{Gm})/E_{Gd} = \rho_G \tag{8}$$

$$(E_B - E_{Bm})/E_{Bd} = \rho_B \tag{9}$$

Here, the specular components (E_{Rm}, E_{Gm}, E_{Bm}) can be separated from (E_R, E_G, E_B), and the diffuse components (E_{Rd}, E_{Gd}, E_{Bd}) can be calculated by the cosine of the incident angle i, where i is the angle between $\boldsymbol{n}$ and $\boldsymbol{s}$ of the feature point.

$$\boldsymbol{\rho}^{'} = \{\rho_R^{'}, \rho_G^{'}, \rho_B^{'}\} = \{\rho_R \times rand_1, \rho_G \times rand_2, \rho_B \times rand_3\} \tag{10}$$

The random color reflectance factors $(\rho_R^{'}, \rho_G^{'}, \rho_B^{'})$ are given using Eq.(10), where $(rand_1, rand_2, rand_3)$ is the randomized real value between 0 and 1. The white color points has the values $(1, 1, 1)$ for reflectance factor.

To estimate color reflectance factor, the training data $(E_{1R}^{'}$, $E_{1G}^{'}$, $E_{1B}^{'}$, $E_{2R}^{'}$, $E_{2G}^{'}$, $E_{2B}^{'}$, $E_{3R}^{'}$, $E_{3G}^{'}$, $E_{3B}^{'}$, $E_{4R}^{'}$, $E_{4G}^{'}$, $E_{4B}^{'})$ are given as the input to the NN, as shown in Fig.1-(b). The random color reflectance factors $(\rho_R^{'}, \rho_G^{'}, \rho_B^{'})$ are given as the output data. After learning of the NN, the input data $(E_{1R}$, E_{1G}, E_{1B}, E_{2R}, E_{2G}, E_{2B}, E_{3R}, E_{3G}, E_{3B}, E_{4R}, E_{4G}, $E_{4B})$ of a test object are given, then, the color reflectance factor (ρ_R, ρ_G, ρ_B) of the test object as the output data is obtained from NN shown in Fig1-(b) in generalization for the test object.

4 Neural Network Based Rendering

Given the geometric shape and the color reflectance factor, a virtual image can be rendered for any viewpoint under any direction of the light source. The rendered image irradiance, In general, E, can be represented using the incident angle i, the emittance angle e, and the phase angle g with the color reflectance factor ρ.

The previous approach of neural network based rendering [5] uses (i, e, g) shown in Eq.(11).

$$\begin{aligned} i &= cos^{-1}(\boldsymbol{n} \cdot \boldsymbol{s}) \\ e &= cos^{-1}(\boldsymbol{n} \cdot \boldsymbol{v}) \\ g &= cos^{-1}(\boldsymbol{v} \cdot \boldsymbol{s}) \end{aligned} \tag{11}$$

Here, instead of using the phase angle g, the angle h is defined in Eq.(14).

$$h = cos^{-1}(\boldsymbol{d} \cdot \boldsymbol{n}) \tag{12}$$

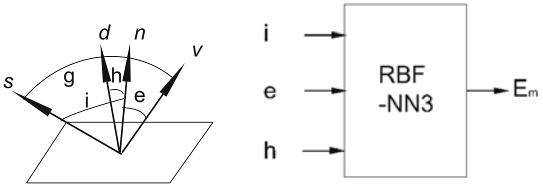
where h is the angle between $\boldsymbol{n}$ and a vector $\boldsymbol{d}$, and the vector $\boldsymbol{d}$ is equally divided vector between $\boldsymbol{s}$ and $\boldsymbol{v}$ as shown in Fig.2-(a). In the proposed approach, E is derived from Eq.(13).

$$E = R_d(\boldsymbol{n}, \boldsymbol{s})\rho + R_m(\boldsymbol{n}, \boldsymbol{s}, \boldsymbol{v}) = E_d(i)\rho + E_m(i, e, h) \tag{13}$$

Here, the range of (i, e, h) is given as

$$0^\circ \leq i \leq 90^\circ, 0^\circ \leq e \leq 90^\circ, 0^\circ \leq h \leq 90^\circ \tag{14}$$

The learning of the mapping of (i, e, h) to E_m learned for the specular components of the test object. After the learning, E_m is generalized using rendering NN. The structure of this RBF-NN is shown in Fig.2-(b). Here, E_d can be calculated by $\cos i = \boldsymbol{n} \cdot \boldsymbol{s}$ and E_m is estimated using an RBF neural network. The rendered image irradiance E can be calculated from Eq.(13).



(a) (i, e, h)

(b) NN Rendering
for Specular Component

Fig. 2. Neural Network Based Rendering

5 Experimental Results

Fig.3 illustrates the observation environment. Four light sources are used to illuminate the test object. Four images are obtained under four different conditions of illumination for each object pose during rotation. The test object is rotated with every 1 degree between 0 to 359 degrees. A total of 360 × 4 images are taken to perform the self-calibration during rotation.

The recovered shape of the target object using self-calibration and NN is shown in Fig.4. Fig.4-(a) linearly encodes the slope angle e (i.e., the angle between the surface normal and the viewing direction) as a gray value in the range of black ($e = 0$) to white ($e = \pi/2$), while Fig.4-(b) plots the aspect angle (i.e., the projection of the surface normal onto the XY plane) as a short line segment. Fig.4-(c) encodes the albedo (color reflectance factor). Both the surface orientation and color reflectance factor are recovered by the proposed self-calibration approach without using any calibration sphere.

Next, the virtual images are generated from the 3-D model aquired by NN. The real image is shown in Fig.5-(a), the virtual image under any light source direction is shown in Fig.5-(b) and Fig.5-(c). It is shown that the virtual image rendering gives the realistic feelings for both of them.

The height distribution obtained by the integration of surface orientation is shown in Fig.6-(a). Rotating this height distribution and the rendering NN can generate a realistic virtual image at any view point. The results are shown in Fig.6-(b) and Fig.6-(c) from the different views.

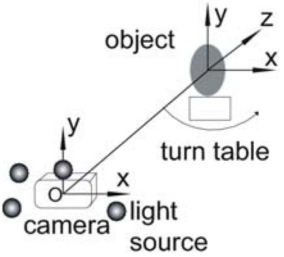
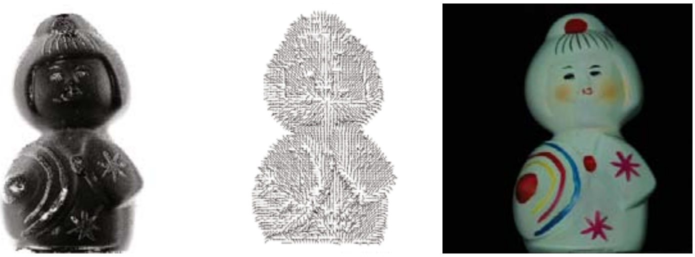
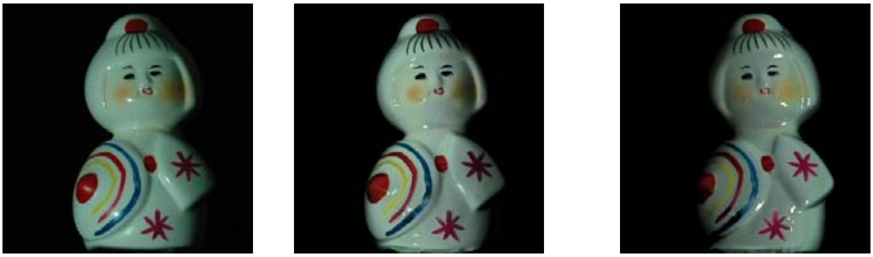


Fig. 3. Observation Environment



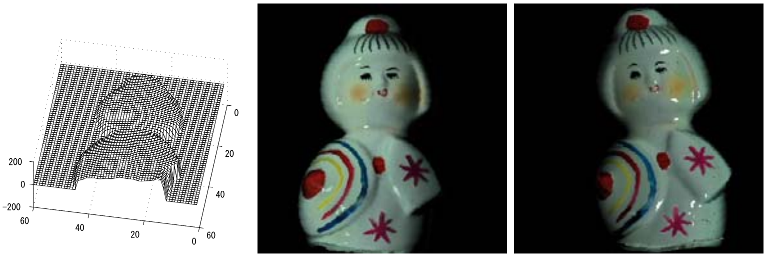
(a) Slope (b) Aspect (c) Color Reflectance

Fig. 4. Recovered Surface Shape and Reflectance



(a) Actual Input Image (b) Any Light Source Direction 1 (c) Any Light Source Direction 2

Fig. 5. Results of Virtual Image Rendering



(a) Height Distribution (b) Any View Point 1 (c) Any View Point 2

Fig. 6. Results of Virtual Image Rendering

6 Conclusion

This paper proposed a new method of self-calibration and color image rendering using RBF-NN without using any calibration object. Instead, the rotation of the test object is used in self-calibration. With four input images, both the surface orientation and color reflectance factor are obtained using NN from the target object itself during rotation. Further, a virtual image under any viewpoint and any direction of light source can be obtained with rendering NN. The proposed approach has an advantage that entire approach is quite empirical without using any calibration object which has the same reflectance property as the target object. Experimental results are shown for a real object. Cast shadow is another problem but this remains as the future work.

Acknowledgements

The authors thank lab. colleagues at NIT and Chubu Univ. for useful discussions. Ding's research is supported by the Hori Information Science Promotion Foundation. Iwahori's research is supported by JSPS Grant-in-Aid for Scientific Research (C) (2050 0168) and Chubu Univ. Grant. Woodham's research is supported by NSERC.

References

1. Woodham, R.J.: Photometric method for determining surface orientation from multiple images. Opt. Engineering, 139–144 (1980)
2. Woodham, R.J.: Gradient and curvature from the photometric-stereo method, including local confidence estimation. J. Opt. Soc. Am., A, 3050–3068 (1994)
3. Iwahori, Y., Woodham, R.J., Bagheri, A.: Principal components analysis and neural network implementation of photometric stereo. In: Proc. IEEE Workshop on Physics-Based Modeling in Computer Vision, pp. 117–125 (1995)
4. Iwahori, Y., Woodham, R.J., Bhuiyan, S., Ishii, N.: Neural Network Based Photometric Stereo for Object with Non-Uniform Reflectance Factor. In: Proceedingsof 6th International Conference on Neural Information Processing (ICONIP 1999), vol. III, pp. 1213–1218 (1999)
5. Kawanaka, H., Iwahori, Y., Woodham, R.J., Funahashi, K.: Color Photometric Stereo and Virtual Image Rendering Using Neural Network. Trans. of IEICE (in Japanese) J89-D-II(2), 381–392 (2006)
6. Iwahori, Y., Watanabe, Y., Funahashi, K., Woodham, R.J.: Self-Calibrated Neural Network Based Photometric Stereo. International Journal of Computer and Information Science 3(1), 40359 (2002)

Similarity Grouping of Paintings by Distance Measure and Self Organizing Map

Naohiro Ishii[1], Yusaku Tokuda[1], Ippei Torii[1], and Tomomi Kanda[2]

[1] Aichi Institute of Technology, Yakusacho, Toyota, Japan
[2] Aichi Prefectural University of Fine Arts and Music, Aichi Pref., Japan
{ishii,mac}@aitech.ac.jp

Abstract. Paintings have some sensibility information to human hearts. It is expected in paintings to process such sensibility information by computers effectively. For appreciation of paintings, grouping of paintings with similar sensitivity will be helpful to visitors as in painting gallery. In this paper, we developed a distance measure to group and classify similar paintings. Further, we applied the self organizing method (SOM) by two layered neural network to classify paintings. Then, the attributes of the sensibility of paintings are checked first. Next, color attributes of paintings are also checked. Paintings data with these attributes were computed by applying these techniques. Relatively well grouped results for the classification of paintings were obtained by the proposed method.

1 Introduction

Much attention has been often paid for in the painting gallery information and painting databases for the enrichment of human living life. In the field of arts, painting will play an important role for giving the strong impression to visitors in the gallery and the museum. In the progress of multimedia processing by computers and networks, painting images are expected to be processed for the semantic information as much as possible effectively. Then, paintings have their respective attributes and characteristics[1,2,5]. Some visitors will have subjective preference to paintings. Then similarity among pictures will be a useful concept in paintings images[3,4]. To offer the painting event information in the gallery or museum, a certain similarity information will be useful to those visitors by having the announcement of new guide of paintings events. In this paper, how to get some similarity information of paintings, is developed. To make clear the similarity painting information, vector representation of paintings is firstly discussed in their relations.

Painting has essentially physical attributes as color features of brightness, chromaticity, saturation, surroundings colors, and color spatial organization. Further, we have some impressions from paintings, which will come from various factors as motif, composition and physical colors. Sensitive impression is important for paintings from which the visitors have their respective ones. In this paper, impressive words are taken for respective painting. By using this impressive words, the similarity of paintings are computed by applying the distance function measure. Similar paintings from

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 713–720, 2009.

the given picture will be near in the distance. Then, pictures with the relation, will make a similar group. Here, Euclidean distance relation was applied first for the grouping o paintings. Since this method takes some steps for the grouping, a well known method of self-organizing map(SOM)[6,7] is applied to compare the group paintings, which was developed by Kohonen [6,7]. To compare the developed method of distance relation here with the SOM method, some experimental computations were carried out for paintings grouping.

2 Impression and Color Words for Expression of Paintings

Expression of paintings is important to classify the similarity grouping of them. Paintings are composed of composition of objects and their color expressions. Human perceives some impressions from paintings. Then, it is a problem how to express impressions from paintings. Here, impression words and color words are discussed to make similar group of paintings.

2.1 Impression Words for Paintings

Paintings often give us emotional or reasonable impressions, which are difficult to express in human perceptional means, exactly. But, impressions are expected to present in some natural words. Here, natural words of impression for paintings are described. For example, painting 1 is expressed in words as night, rich, warm, soft, darkness, light, cool. Paintings 2 is expressed in words as fresh, pleasure, rhythmical, dense, perplexity, lust, suffering, moment, sedition, anger, lie. Painting 3 is expressed in words as heat, move, dry, earth, desolately, primitiveness, rejoicing. Painting 4 is expressed in words as clearly, lazy, man and woman, young, healthy, desolately, et al. Other paintings are also expressed in words. These words are considered as attributes of paintings which are components of coordinate system. When an attribute exists in the painting, the corresponding component becomes 1, otherwise 0. Thus, the painting is considered as a corresponding vector. So, we call here painting vector whose components are 1 or 0 corresponding to each attribute, respectively.

2.2 Color Words for Paintings

Paintings are constructed of composition of objects and colors in their respective regions. In their respective scene, we consider colors and brightness.Then, color will be an important factor for the impression of paintings. The visual property corresponding to the categories called red, blue , yellow and others. These colors have properties of hue, saturation and brightness. Here, we simpy apply color naming words as red, blue, yellow, black, white, gray, pink, brown, green, et al. These are sub-components of color vector expression in physical expression of paintings.

Brightness properties also are sub-components of vectors in physical expression of paintings.

3 Vector Relations among Paintings

Similarity of paintings is developed here by the relations of vector of paintings. A painting is represented a vector notation as shown in the previous section 2. Each component of the vector shows an attribute of the painting. Then, similarity of two paintings will mean by existence of same attributes between them. Two paintings will be more similar if the same attributes are more counted between them. Similarity of paintings is defined by using ε distance($\varepsilon > 0$) in the following.

Let the vector notation of two paintings be A and B . Two paintings A and B are similar in ε distance denoted by absolute value , when the equation (1) holds,

$$\|A - B\| \le \varepsilon \tag{1}$$

which is denoted by $A \approx B$ in ε distance. Equation (1) is shown as in Fig.1.

Fig. 1. Distance between A and B

In the relation of vectors, $A \approx B$, the transitive relation does not necessarily hold, i.e, When $A \approx B$ and $B \approx C$ hold, $A \approx C$ does not necessarily hold. To make similar paintings group, relations among three vectors A , B and C of paintings are discussed. Three vectors A , B and C of paintings will be classified into four classes by using equation (1).

Type class [1]

Among three vectors A , B and C of paintings, the following equation holds,

$$\|A - B\| \le \varepsilon \ , \ \|B - C\| \le \varepsilon \ \text{ and } \ \|A - C\| \le \varepsilon \tag{2}$$

This class is schematically described as shown in Fig.2. From equation(2), $\|A - B\| \le \varepsilon$ and $\|A - C\| \le \varepsilon$ show that vector A (picture A) is similar to vectors B (picture B) and C (picture C) in the sense of ε distance. At the same time, vector B is similar to vectors A and C , while vector C is similar to vectors A and B in the sense of ε distance. This class is relatively restricted, since three equations among vectors must be held. Next class is relaxed in ε equations.

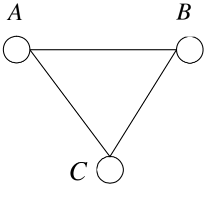


Fig. 2. Schematic diagram of type class [1]

From equation (2), $\|A-B\| \leq \varepsilon$ and $\|A-C\| \leq \varepsilon$ show that vector A (picture A)is similar to vectors B (picture B) and C (picture C) in the sense of ε distance. At the same time, vector B is similar to vectors A and C , while vector C is similar to vectors A and B in the sense of ε distance. This class is relatively restricted, since three equations among vectors must be held. Next class is relaxed in ε equations.

Type class [2]

Among three vectors A , B and C of paintings, the following equation holds,

$$\|A-B\| \leq \varepsilon \ , \ \|A-C\| \leq \varepsilon \ \text{and} \ \|B-C\| \leq \varepsilon \tag{3}$$

From equation(3), $\|A-B\| \leq \varepsilon$ and $\|A-C\| \leq \varepsilon$ show that vector A (picture A) is similar to vectors B (picture B) and C (picture C) in the sense of ε distance. But, vector B is not similar to vector C in the sense of ε distance, which shows vector B is apart from C more than ε in the distance. This class is schematically described in Fig.3.

$$\|A-B\| \leq \varepsilon \ , \ \|B-C\| \leq \varepsilon \ \text{and} \ \|A-C\| \leq \varepsilon . \tag{4}$$

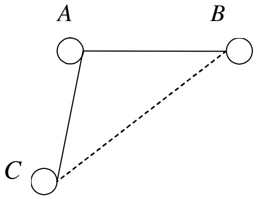


Fig. 3. Schematic diagram of type class [2]

Further, the class[2] in equation (3) is also same in the class with the following equations,

$$\|A-C\| \leq \varepsilon \ , \ \|B-C\| \leq \varepsilon \ \text{and} \ \|A-B\| \leq \varepsilon \tag{5}$$

Type class [3]

Among three vectors A , B and C of paintings, the following equation holds,

$$\|A-C\| \leq \varepsilon \ , \ \|A-B\| \leq \varepsilon \ \text{and} \ \|B-C\| \leq \varepsilon \tag{6}$$

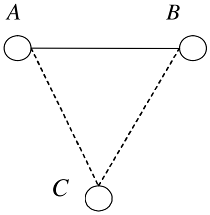


Fig. 4. Schematic diagram of type class [3]

Type class [4]

Among three vectors A , B and C of paintings, the following equation holds,

$$\|A-B\| \leq \varepsilon \ , \ \|A-C\| \leq \varepsilon \ \text{and} \ \|B-C\| \leq \varepsilon \tag{7}$$

Grouping of similar paintings with near distance from A is ordered as Type class[1], Type class[2] ,Type class[3] and Type class[4] in order. From the vector A, the distances are evaluated in the respective type class. In the type class[1], the distance of the vector B or C from the vector A , is $\leq \varepsilon$. The distance between B and C is $\leq \varepsilon$. In the type class[2], the distance of the vector B or C from the vector A is $\leq \varepsilon$, while the distance between B and C is $\leq 2\varepsilon$, since $\left\|B-C\right\| = \left\|(A-B)+(B-C)\right\| \leq \left\|A-B\right\| + \left\|B-C\right\| = 2\varepsilon$. In the type class[3], only the distance of the vector B from the vector A , is $\leq \varepsilon$. Similarity problem in paintings is described here to count how many type classes [1], [2] and [3] in $\leq \varepsilon$ distance from the started vector A. Much counts of type classes [1] and [2] show the similarity will be high in these local paintings in $\leq \varepsilon$ distance. To evaluate further similarity of paintings, next search and comparison are needed. The above searching steps are summarized as follows,

* (1) Nearest vectors(the first near pictures) from the object vector A (object picture) are searched first in the sense of ε equation (1).

* (2) Second, vectors $\{x\}$ (the second near pictures) from the object vector A are searched , which satisfy the inequality $\varepsilon < x \leq 2\varepsilon$.

* (3) Inequality $m\varepsilon < x \leq (m+1)\varepsilon$ where $m \geq 3$, is iterated to find vectors from vector A. The near vectors * (1) and * (2) are gathered, first from ones with attributes having '1' value of the object vector A. So, the candidate vectors near A is chosen, easily, whose relation is checked in the above conditions.

4 Experimental Results of Similarity of Paintings

Similarity of paintings are computed as an example. An objective painting is given. The paintings shown in Fig.5, are western ones in art museums in Europe. First, similar (near) paintings to the 20 are searched by * (1) and * (2) in the above. Assume here the number 20 in Fig.6 is the objective painting. The searched paintings near to the 20 by steps * (1) and * (2), are shown on the middle row in Fig. 5. Here, the art of 20 is named as ' Breakfast', The 40 is 'Painter's daughters chasing a butterfly'. The 1 is 'Woman making her toilet'. The 27 is 'Sacred and profane love'. The 11 is 'Philosopher meditating'. The 41 is 'Last supper'. The 9 is 'St. Andrew'. The 20, ' Breakfast', is given in advance. On the bottom row in Fig.5, paintings far from 20, are shown. The 40 was near to the 20 in Fig.5. Next, what kind of paintings are similar to the 40 picture in sensibility? To answer this question, the searching steps * (1) and * (2) were carried out. Then, by steps * (1) and * (2), similar paintings to 40 are searched as shown on the middle row in Fig.6. Thus, the paintings computed near from 20 to 40, are almost similar by comparing top and middle rows shown in Fig.5 and Fig.6.

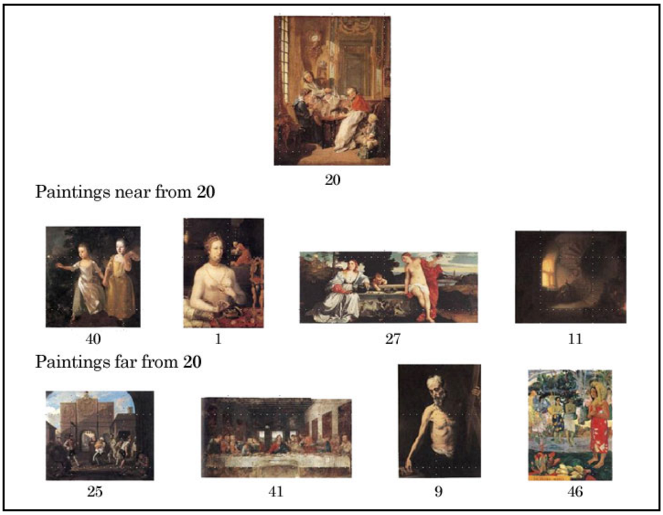


Fig. 5. Paintings near from and far from an objective picture, 20 in sensibility

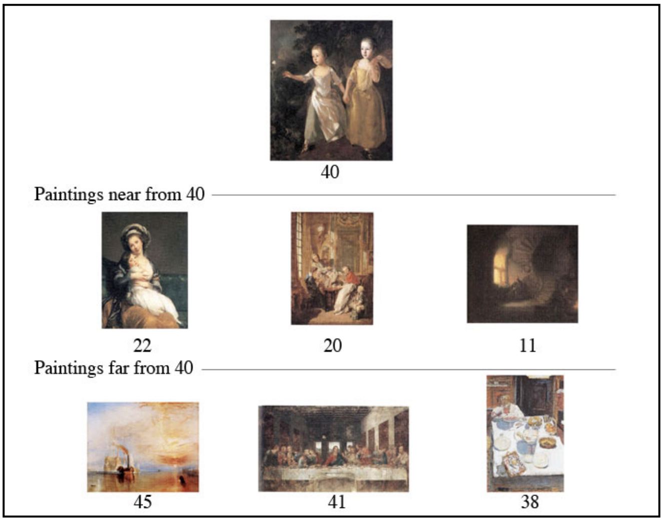


Fig. 6. Paintings near from and far from an objective picture, 40 in sensibility

5 Top Down Approach by Self Organizing Map

The low-dimensional ,ordered representation of data with high-dimensional data, isrealized by self organizing map(SOM) proposed by Kohonen(1989)[6,7].

The self organizing map process can be described in the following mathematical form. The input consists of n-dimensional data vectors, each of the form is as follows,

$$x = (x_1, x_2, ..., x_n) \tag{8}$$

The input x is inputted to the two layered neural network which consists of the inputted layer and the output layer(mapping layer). In the mapping layer, the distance between the weight vector w of each node neuron and the input vector x , is computed as follows,

$$d_m = \sqrt{\sum_{i=1}^{n} (x_i - w_{mi})^2} \tag{9}$$

where suffix m in equation(9) means the m -th node neuron on the mapping layer. The minimum distance of the node neuron is chosen as the winner node. The weight vector w is modified according to equation (10).

$$\triangle w_{mi} = \eta h(m, m^*)(x_i - w_{mi}) \tag{10}$$

where $h(m, m^*)$ is called neighborhood function as shown in (11) and η is a positive constant..

$$h(m, m^*) = \exp(-|m - m^*|^2 / \sigma^2) \tag{11}$$

The goal of the training process is to determine the n-dimensional weight vectors of nodes(neurons). Experimental results by the SOM method are shown in Fig.7, which are grouped as the Volonoi diagram. The number in Fig.7 shows painting number for experiments as shown in Figs. 5 and 6. The bold alphabet A, B, C and D

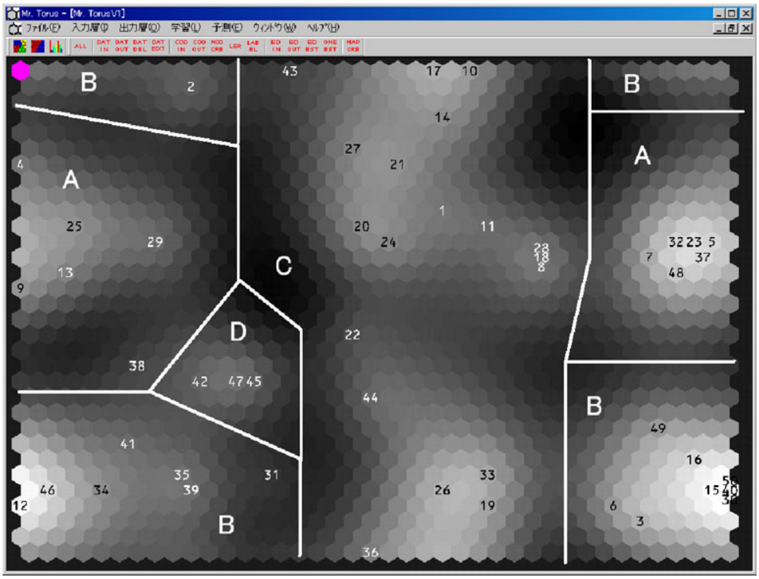


Fig. 7. Similarity grouping by SOM method in the Voronoi diagram

show the grouping made by the SOM method, which are blocked by white lines. In the same C group, black and white numbers will make different sub-groups. Thus, similarity grouping visualization of paintings will be useful by the SOM representation from high dimensional impression words. By comparing Fig. 7 made by the SOM method with Fig. 5 and Fig. 6 made by near distance method proposed here, there are almost same results and a slight different interpretation, which will be needed to do further analysis.

6 Conclusion

Similar grouping of paintings are carried out by near distance method proposed here, which is a step by step seeking method of near paintings. This method is a bottom-up method to make classification of similar paintings, which is expected to make clear the near relations of paintings. Then, paintings are classified by this proposed method. As the top-down method, the SOM method was applied to classify and group the similar paintings, which is useful for the reduction of dimensions of data and low dimensional visualization. The proposed method for paintings classification is compared with the SOM method.

References

1. Yelizaveta, M., Tat-Seng, C., Irina, A.: Analysis and Retrieval of Paintings Using Artistic Color Concepts. In: Proc. IEEE Int. Conference on Multimedia and Expo 2005, pp. 1246–1249. IEEE Pub., Los Alamitos (2005)
2. Kahol, K., French, J., Bratton, L., Panchannathan, S.: Learning and Perceiving Colors Haptically. In: Proc. ASSETS 2006, pp. 173–180. ACM Press, New York (2006)
3. Rigau, J., Feixas, M., Sbert, M.: Informational Aesthetics Measures. In: IEEE Computer Graphics and Applications, pp. 24–34. IEEE Comp. Society, Los Alamitos (2008)
4. Li, M., et al.: The Similarity Metric. IEEE Trans. Information Theory 50(12), 3250–3264 (2004)
5. Kobayashi, M., Yoshiki, K.: Mosaic Image of Dominant Color Useful for Analysis of Color in Painting Arts. In: Proc. of AIC Colo 2005, pp. 627–630 (2005)
6. Lefebvre, G., Lautent, C., Ros, J., Garcia, C.: Supervised Image Classification by SOM Activity Map Comparison. In: Proc. 10th Int. Conference on Pattern Recognition (ICPR 2006), pp. 728–731. IEEE Computer Society Press, Los Alamitos (2006)
7. Korting, T.S., Fonseca, L.M.G.: Expectation-Maximization×Self-Organizing Maps for Image Classification. In: Proc. IEEE Int. Conference on Signal Image Technology and Internet Based Systems, pp. 359–364 (2008)

Localization in Wireless Sensor Networks by Fuzzy Logic System

Shu-Yin Chiang and Jin-Long Wang

Department of Information and Telecommunications Engineering,
Ming Chuan University
5 Deh-Ming Road, Gwei-Shan, Taoyuan 333, Taiwan
sychiang@mail.mcu.edu.tw, jlwang@mcu.edu.tw

Abstract. This paper presents a novel algorithm for localization in wireless sensor networks utilizing a fuzzy inference system at each sensor node. The algorithm using fuzzy distance measuring based on received signal strength information (RSS). The advantage of employing the RSS information is that no extra hardware is needed for localization. The simulation results and indoor experiments demonstrate that the proposed scheme employing fuzzy logic system can localize the mobile sensor nodes with certain accuracy.

Keywords: wireless sensor networks (WSN), fuzzy logic systems, localization, received signal strength (RSS).

1 Introduction

In recent years, with the rapid advances in the hardware implementation, the deployment of large numbers of low cost wireless sensors is a technique for many applications, such as target tracking, intruder detection, animal monitoring and real-time traffic monitoring. Localization in individual mobile nodes is a very important requirement for the successful operation of mobile ad-hoc autonomous sensor networks. Sensor network localization algorithms estimate the locations of sensors with initially unknown location information by using knowledge of the absolute positions of a few sensors and inter-sensor measurements such as distance and bearing measurements. Measurement techniques in WSN localization can be broadly classified into two categories: range-based localization methods and range-free localization methods [7]. The range-based localization needs hardware to obtain distances or angles by measuring the time of arrival (TOA), time difference of arrival (TDOA) and angle of arrival (AOA) of signals. In other words, range-based localization methods depend on the additional hardware such as antennas to distinguish the angles of arrivals etc. In contrast, the range-free localization method can be implemented by measuring the received signal strength indicator (RSS) [3]. The range-free localization schemes provide simpler and more economic than range-based ones. Therefore, this paper will focus on the range-free localization scheme based on the RSS information.

In this work, to enhance estimation accuracy, we propose a scheme based on Fuzzy distance measurement which considers multiple parameters: the received signal

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 721–728, 2009.

strength, the distance between the anchor node and mobile node, the moving direction, and the previous location. Through the fuzzy logic system, the weighting factor is used to fit the localization process.

The remaining paper is organized as follows. Section 2, describes some related works. Section 3 states the fuzzy logic algorithm for localization. In Section 4, we discuss the performance measurements and simulation results and finally in Section 5, the conclusions are stated.

2 Related Work

Extensive research has been done on localization for wireless sensor networks. In localization without complicated hardware design, the method may be implemented on the pre-knowledge of location of the reference nodes. Then applying the information of the received signal strength information, the sensor nodes can be allocated. The related works about this paper are discussed in the followings.

2.1 Received Powers

Since this work focuses on the RSS for localization then the relationship between the distance and the signal power of the receiver is formulated as follows ([1]):

$$P_r(d) = P(d_0) - 10n\log(\frac{d}{d_0}), \tag{1}$$

where $P_r(d)$ is the receiving signal power of receiver when the transmitter and receiver distance is *d*, and $P_r(d_0)$ is the power when *d* is reference distance d_0. Hence, the estimated distance *d* can be obtained by

$$d = d_0 \times 10^{\frac{P_r(d_0)-P_r(d)}{10n}}. \tag{2}$$

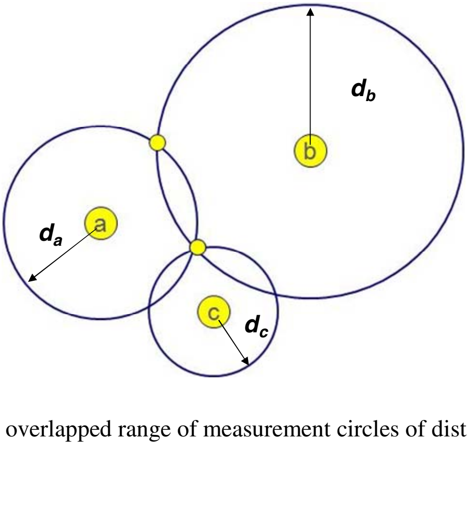


Fig. 1. Senor lies in the overlapped range of measurement circles of distance d_a, d_b, and d_c

Therefore, according to the triangular positioning method, the location of the sensor node lies in the range circles and the overlapped region as shown in Figure 1.

2.2 Weighted Centroid Algorithm

Use the centroid localization method by using the edge weights of adjacent reference nodes based on the fuzzy model with the weighting factor according to the received signal strength information. In [2], they propose fuzzy model to compute edge weights $\tilde{y}$, which combines partial information in each RSS value is computed by the weighted average of the y^l, where the output of the fuzzy rule.

$$\tilde{y} = \frac{\sum_{l=1}^{M} y^l w^l}{\sum_{l=1}^{M} w^l} \tag{3}$$

After calculating the edge weights, the weighted centroid algorithm estimates the node position by the following procedure. Let the position of adjacent reference nodes are (X_1,Y_1), (X_2,Y_2), …, (X_n,Y_n), respectively. Then the estimated position calculates as following weighted centroid formula:

$$(X_{est},Y_{est}) = \left[\frac{\tilde{y}_1 \cdot X_1 + \ldots + \tilde{y}_n \cdot X_n}{\sum_{i=1}^{N} \tilde{y}_i}, \frac{\tilde{y}_1 \cdot Y_1 + \ldots + \tilde{y}_n \cdot Y_n}{\sum_{i=1}^{N} \tilde{y}_i} \right], \tag{4}$$

for $i=1,2,\ldots,N$, where N is the number of adjacent reference nodes. Then combining the above related work, we have the proposed algorithm for sensor localization in the next section.

3 Fuzzy Localization Scheme

The proposed scheme involves fuzzy system which approximates any nonlinear function to arbitrary accuracy with only a small number of fuzzy rules [6]. Those parameters are RSS and distance. Assume the direction is a given value. First, let us define the membership function for the received signal strength RSS_i, $f(RSS_i)$, from neighboring anchor i^{th} node. Second, the membership function of the distance d_i, $f(d_i)$, between the sensor node and the neighboring anchor i^{th} node. The distance d_i is obtained by equation (2). The corresponding membership functions, $f(RSS_i)$and $f(d_i)$ are shown in Figures 2 and 3, respectively. In Figures 2 and 3, the parameters set (α, β, γ) and (A, B, C) are adjustable values according the environmental situation and their units are dBm and meter, respectively.

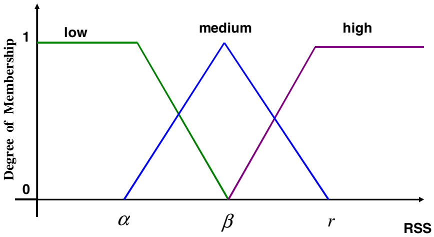


Fig. 2. Membership Function of *RSS*

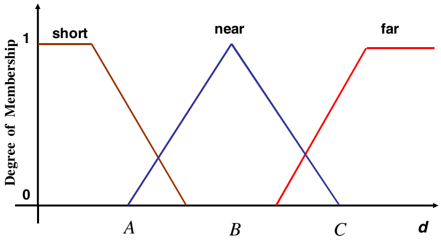


Fig. 3. Membership Function of *d*

The weighting factor for evaluation the criteria for the membership functions defined above is used according to equation (3). The weighting criteria is that the farer the distance the less weighting for each sensor node. The Fuzzy Logic System uses the Tagagi-Sugeno-Kang (TSK) model [9] and the inputs are *RSS* and *d*. The rule is as follows:

$$\text{If the } RSS_i \text{ is in } X \text{ and the distance } d_i \text{ is in } Y \text{ then } Z_i = min\ \{\ a_j,\ d_i\ \}, \tag{5}$$

where (*X*, *Y*) in set pair as { (high, short), (medium, near), (low, far) } of Figures 2 and 3, and a_j is the obtained distance with the parameters of RSS_i, through the equation (2). Therefore, Z_i is the estimated distance with the TSK model shown in Figure 4.

Next, define the weighting w_i, according to the rule: the shorter the distance, the higher weighting factor to locate the node position. After the estimated distance obtained, then by the given direction and previous location information, we have the estimated location of the mobile sensor node.

$$(X_{new}, Y_{new}) = D \times (X_{est}, Y_{est}) + (X_{pre}, Y_{pre}), \tag{6}$$

where (X_{est}, Y_{est}) is the estimated location, *D* is the moving direction and (X_{pre}, Y_{pre}) is the previous location of the mobile senor node.

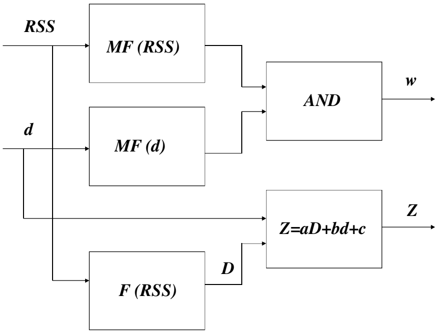


Fig. 4. TSK model for Fuzzy Logic System

4 Performance Evaluations

The system is simulated with MATLAB 7 version for the proposed scheme. By generating the corresponding *RSS* and estimated d as inputs, the fuzzy logic system is used to get the distance as output with weighting w_i, we have the simulation percentage error is about 7% shown in Figure 5.

Then indoor experiments are based on the hardware: Tmote Sky wireless sensor node [8] (the hardware structure shown in Figure 6), which works with CC2420 radio chip at frequency 2.4GHz by IEEE 802.15.4 ZigBee protocol and wireless transceiver with data rate 250kbps. The relationship between receiving signal strength power and RSSI register value is shown in Figure 7. Therefore, the register value needs to be mapped to RF power value in dBm scale first, then applying the equation (2) and (3)to get the distance.

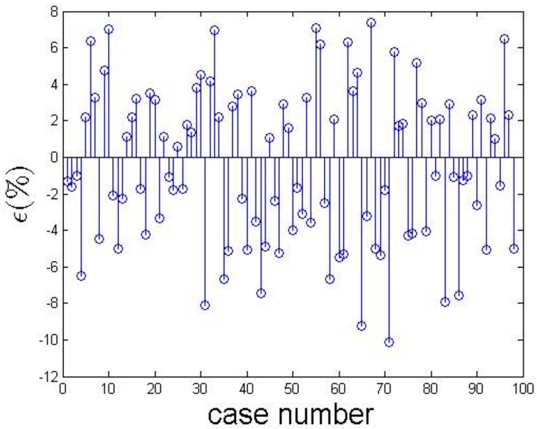


Fig. 5. Accuracy of the simulation with the proposed scheme

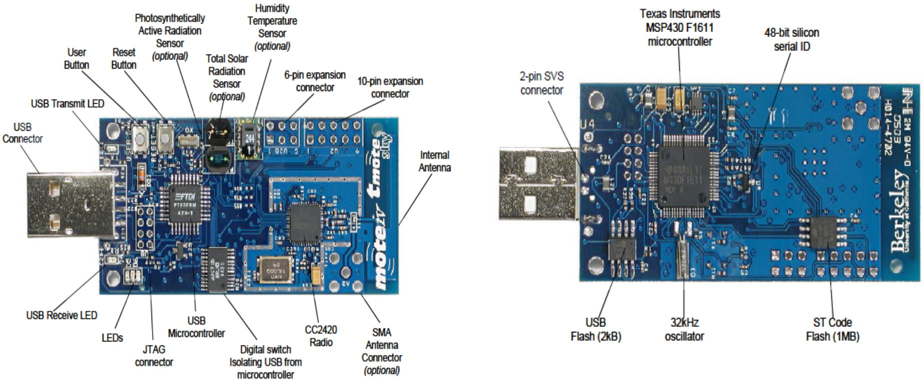

Photosynthetically Active Radiation Sensor (optional)
Humidity Temperature Sensor (optional)
User Button
Reset Button
Total Solar Radiation Sensor (optional)
6-pin expansion connector
10-pin expansion connector
USB Transmit LED
USB Connector
Internal Antenna
USB Receive LED
LEDs
JTAG connector
USB Microcontroller
Digital switch Isolating USB from microcontroller
CC2420 Radio
SMA Antenna Connector (optional)
Texas Instruments MSP430 F1611 microcontroller
48-bit silicon serial ID
2-pin SVS connector
USB Flash (2kB)
32kHz oscillator
ST Code Flash (1MB)

Fig. 6. Front and Back of Tmote Sky module

The performance of the location estimation algorithm by fuzzy logic system is examined. First, we set up the experiment in an indoor gym without obstacles on the floor and the floor is gridded with 2 meters apart and the total working space is 100m^2 as shown in Figure 8 [4]. Next put the anchors node on the floor as the blue nodes in Figure 7. Finally, put the mobile sensor node on the grid point one by one to estimate its location by the above fuzzy logic system with suitable parameters (α, β, γ) and (A, B, C) and the numerical quantities are expressed in dBm and meters, respectively. The system runs 100 times indoor experiment by putting the sensor node on different grid point of Figure 8. Simulations are conducted to investigate how the average location error is by different locations. Due to the *RSS* is affected easily by the environment, the indoor experiment accuracy of the result is shown in Figure 9. From this figure, we have the average error about 16%.

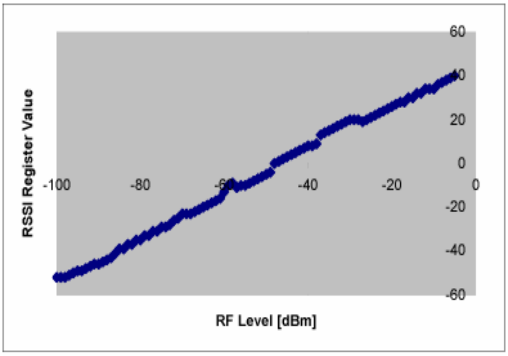


Fig. 7. Received Signal Strength Indicator mapping to RF Power in dBm unit

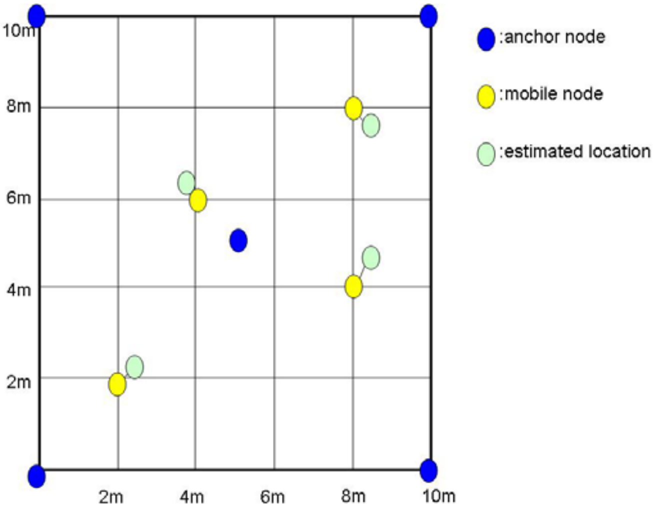


Fig. 8. Indoor Experiment nodes locations and grid-points

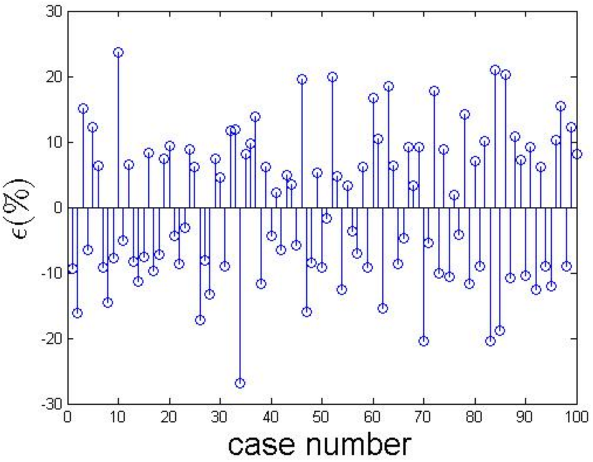


Fig. 9. Accuracy of the indoor experiment with the proposed scheme

5 Conclusions

In this paper, the fuzzy based localization scheme for wireless sensor networks (WSN) is presented. In our proposed method, the sensor nodes do not need any complicated hardware to obtain the information for localization. The sensor nodes positions are estimated through fuzzy logic system with the RSSI between the sensor node itself and its neighbor anchor nodes. The proposed method is applied to both computer simulation and indoor experiments. Both results show the performance with certain accuracy. Therefore, the proposed scheme can be applied to more complicated sensor network systems and implement the sensor localization.

References

1. Feng, X., Gao, Z., Yang, M., Xiong, S.: Fuzzy Distance Measuring Based on RSSI in Wireless Sensor Network. In: IEEE Proceedings of 3rd International Conference on Intelligent System and Knowledge Engineering, pp. 395–400 (2008)
2. Yuna, S., Leea, J., Chunga, W., Kima, E., Kimb, S.: A Soft Computing Approach to Localization in Wireless Sensor Networks. Expert Systems with Applications 36(4), 7552–7561 (2009)
3. Awad, A., Frunzke, T., Dressler, F.: Adaptive Distance Estimation and Localization in WSN using RSSI Measures. In: 10th Euromicro Conference on Digital System Design Architectures, pp. 471–478 (2007)
4. Dharne, A.G., Lee, J., Jayasuriya, S.: Using Fuzzy Logic for Localization in Mobile Sensor Networks: Simulations and Experiments. In: IEEE Proceedings of the American Control Conference, pp. 2066–2072 (2006)
5. Mao, G., Fidan, B., Anderson, B.: Wireless sensor network localization techniques. Computer Networks 51(10), 2529–2553 (2007)

6. Lee, J., Yoo, S.-J., Lee, D.C.: Fuzzy Logic Adaptive Mobile Location Estimation. In: International Federation for Information Processing, pp. 626–634 (2004)
7. Wann, C.-D., Chin, H.-C.: Hybrid TOA/RSSI Wireless Location with Unconstrained Nonlinear Optimization for Indoor UWB Channels. In: IEEE proceedings for WCNC, pp. 3943–3948 (2007)
8. Tmote Sky, `http://www.moteiv.com`
9. Schnitman, L., Felippe de Souza, J.A.M., Yoneyama, T.: Takagi-Sugeno-Kang Fuzzy Structures in Dynamic System Modeling. In: Proceedings of the IASTED International Conference on Control and Application, Banff, Canada, pp. 160–165 (2001)

Dynamic Handover Scheme for WiMAX

Jin-Long Wang and Shu-Yin Chiang

Department of Information and Telecommunications Engineering,
Ming Chuan University, Taipei 11103, Taiwan
jlwang@mcu.edu.tw, sychiang@mail.mcu.edu.tw

Abstract. WiMAX is an emerging technology based on the 802.16 standards to provide high speed and broadband wireless access for mobile stations. Handover is a key operation influencing the quality of communication services. In this paper, the scheme of choosing the suitable base station with the best service for a mobile station in a WiMAX network is studied. A new scheme based on fuzzy logic is proposed to employ the important traffic criteria, including bandwidth, dropping rate, blocking rate, and signal strength. Finally, the simulation is used to investigate the performance of proposed scheme. The simulation results show that the proposed schemes have better performance than conventional schemes, and can achieve the higher bandwidth utilization, the lower blocking rate for new calls, and the lower dropping rate for handover calls.

Keywords: WiMAX, Fuzzy logic.

1 Introduction

Based on the IEEE 802.16 standards, WiMAX (Worldwide Interoperability Microwave Access) provides the high bandwidth, broadband wireless access, and continuous data transmission for the stations with high speed mobility. The transmission speed offered by WiMAX is up to 70 Mbps, which approaches the quality of service as the wire access network. There are two major standards supporting WiMAX, including IEEE 802.16-2004 and IEEE 802.16e. The basic physical features of these two standards, such as frequency ranges, distance, and speed, are shown in Table 1 for NLOS (Non-Line Of Sight) and LOS (Line Of Sight). A WiMAX network is composed of base stations (BS) and mobile stations (MS). The base station plays the role of the gateway between the base stations and the wire access networks, while the mobile station represents the end-site from which users can access networks.

Handover is a key operation influencing the quality of communication services. It is necessary to perform the handover operation when the signal strength of the current cell is too weak to support normal communication. Otherwise, the on-going communication will be forced to be terminated as soon as the signal is not strong enough. Namely, the handover provides the continuity of communication when a mobile station moves from one cell to another cell. Moreover, a suitable handover scheme is able to balance the channel allocation and offer better quality to meet the requirements of mobile station.

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 729–735, 2009.

Table 1. Basic physical features of NLOS and LOS

	NLOS	LOS
Frequency ranges	2-11 GHz	10-66 GHz
Distance	8 km	50 km
Speed	70 Mbps	10 Mbps

IEEE 802.16e defines three kinds of handover mechanisms, including Hard Handover, Macro Diversity Handover, and Fast Base Station Switching [1-2, 7]. Hard Handover is a relatively simple mechanism. Using this mechanism, one station can only connect to a base station (BS) at a time. When the signal strength of current base station is lower than the neighbor base station, the mobile station will break the current connection, and reconnect to the neighbor base station with stronger signal strength. Obviously, the Hard Handover will cause a short period of disconnection. Thus, the Hard Handover usually is used for data, but not for the real-time or streaming applications.

By using Macro Diversity Handover, one mobile station can connect more than one base station, which is called "Active Base Station" (ABS). The collection of active base stations is called "Active Set", which is maintained by the mobile station and the base station via MAC (Medium Access Control) management messages. Among the Active Set, one base station is selected by the mobile station as the "Anchor BS" for performing synchronization, registration, and monitoring the control information. In this mechanism, mobile station is able to receive data from more than one base station, and can also send data to several base stations.

Similarly, in Fast Base Station Switching, one mobile station can connect more than one base station. The definition of Active Base Station, Active Set, and Anchor Base Station are the same as Macro Diversity Handover. However, in Fast Base Station Switching, one mobile station can only perform all communication operations with the anchor base station, including uplink, downlink, and management. Especially, the mobile station can change the anchor station for transmitting different data frames. Each data frame can be transmitted to different base station.

In this paper, the fuzzy logic is adopted for handover decision for the WiMAX networks with the handover mechanism of Fast Base Station Switching. The proposed scheme is based on fuzzy logic to provide efficient handover in order to reduce the blocking rate and the dropping rate, and to enhance the system utilization and the communication quality. The important traffic criteria adopted by the fuzzy logic contains the bandwidth utilization, the dropping rate, the blocking rate, and the signal strength.

When a mobile user first enters the WiMAX, the fuzzy logic based handoff scheme is used to choose the suitable anchor BS for accessing. Whenever the signal strength is below the acceptable level, the proposed scheme is used again to take the handover into consideration.

The remaining paper is organized as follows: In section 2, we will describe the system models and previous works. The proposed scheme is described in details in section 3. The simulation model and results are then presented in section 4. Finally, some conclusions are given in section 5.

2 System Models and Previous Works

Handoff decision algorithm includes conventional power level based handoff, user population based handoff, and bandwidth based handoff [3-6]. Those items and the system model will be discussed next.

2.1 System Model

In Mobile WiMAX of IEEE 802.16e, the physical layer adopts SOFDMA (Scalable Orthogonal Frequency Division Multiplexing), and the system model is shown in Figure 1. There are three kinds of base stations, including ABS (Active Base Station), NBS (Neighboring Base Station), and AnBS (Anchor Base Station). The active set of each mobile station is consisting of all base stations that can offer normal uplink and downlink traffic, while the neighbor set is composed of the base stations that can limitedly communicate with mobile station but the signal strength is not sufficient enough to support normal transmission.

2.2 Signal Strength Based Handoff (SSH)

The signal strength received from the candidate wireless network is the only evaluated item, and the base station with the highest power level is selected for handover. Since the network does not take into account of the loading for the base station, the scheme results in non-uniform utilization of system resource and poor QoS.

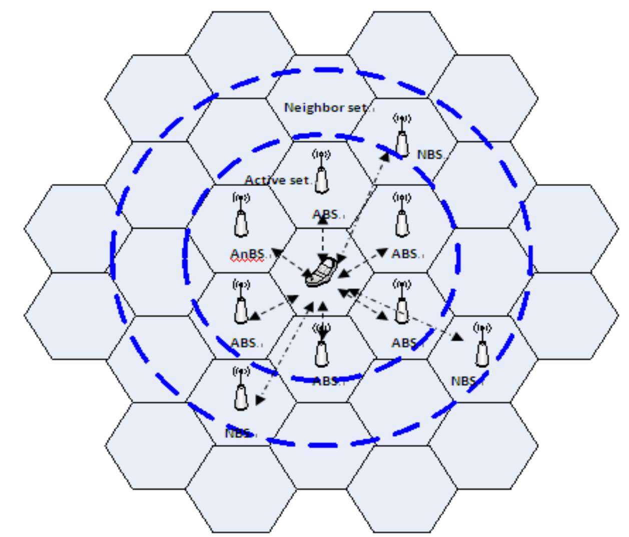


Fig. 1. WiMAX networks

2.3 User Population Based Handover (UPH)

The mobile terminal is handed over to the base station which has its signal strength above the threshold with the minimum number of users. It then remains with this base station until the received signal strength is below the threshold value.

2.4 Bandwidth Based Handover (BWH)

The base station with the minimum account of used bandwidth is selected among the candidate base stations which have received signal strength above the threshold required for normal operation. As in the UPH scheme above, the wireless then stays with the base station until the received signal strength falls below the threshold.

3 Proposed Scheme

In this section, an adaptive fuzzy logic based handover scheme is proposed for the mobile station roaming in the WiMAX networks. There are three steps in this scheme, including Fuzzification, Rule Evaluation, and Defuzzization.

3.1 Fuzzification

In the fuzzification, the input parameters, signal strength, available channels, bandwidth utilization and dropped rate are fuzzified using pre-defined input membership functions. For new calls, three input parameters, including signal strength, available channels and bandwidth utilization, are taken into account. However, for handover calls, three input parameters are the same, except that the bandwidth utilization is replaced by the call dropped rate.

The parameters in the WiMAX network are fed into a fuzzifier, which will transform the real-time measurements into fuzzy sets. The membership values are obtained by mapping the values onto a membership function.

3.2 Rule Evaluation

In the second step, the fuzzified input values are used to evaluate rules for obtaining Fuzzy Decision (FD). The decision set could be classified into four different sets: Yes (Y), Probably Yes (PY), Probably No (PN) and No (N). For example, **IF** the available channels is LowCH, the signal strength is LowSS and the bandwidth utilization is HighU, **THEN** handoff decision=N. Following this, a set of different handoff decisions can be obtained.

After the evaluation of all rules, some FDs have more than one value for the degree of membership. In such situation, the simulation has considered as using the maximum of the membership-degrees.

$$\mathrm{FD}(p) = \max(\mathrm{FN}_1, \mathrm{FN}_2, \ldots, \mathrm{FN}_k), \tag{1}$$

where p is one of four possibilities, including Y, PY, PN, or N and k is the number of input parameters.

3.3 Defuzzification

In the last step, the resultant fuzzy decision sets will be converted to a precise quantity. By using these weighting values and the degree of membership for the FD output, the crisp value of FD is determined by using the following formula:

$$FD = \frac{\sum_{i=1}^{k} M_i \times W_i}{\sum_{i=1}^{k} M_i}, \tag{2}$$

where M_i is the degree of membership in output singleton *i*, and W_i is the FD weighting value for the output singleton *i*. The crisp value of FD can be calculated for each base station. Finally, the base station with the highest FD value will be selected for new calls or handoff calls.

4 Simulation Results

The simulation results of proposed handover scheme are illustrated in this section. In the simulation environment, every base station is equipped with 32 channels. The mobile station based on the proposed handover scheme in the WiMAX network will select one base station and handover to the suitable base station for each evaluation period. Some important performance factors are investigated, including the blocking rate, the dropping rate, and the bandwidth utilization.

4.1 Investigations of Four Handover Schemes

In this subsection, the performance of proposed handover scheme is compared with the conventional three schemes, including the signal strength scheme (SSH), the user population scheme (UPH), and the bandwidth-based scheme (BWH). The blocking

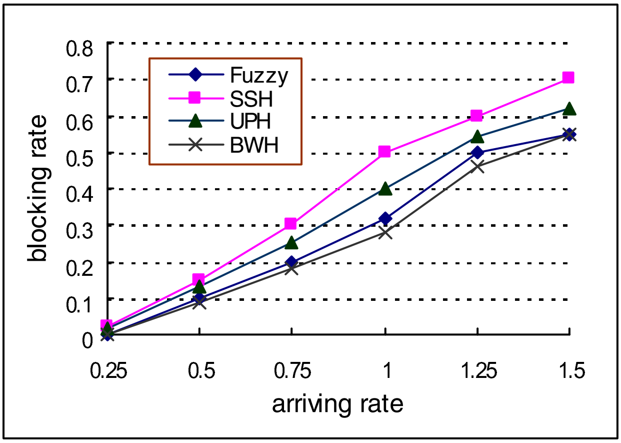


Fig. 2. Blocking rate vs. mean arrival time

rate versus the arrival time is shown in Figure 2. It is obvious that proposed fuzzy handover scheme outperforms the other three schemes because this scheme always selects the base station with the largest number of channels for mobile station. Signal strength based handover scheme (SSH) has the worst because this scheme may select the base station with the strongest signal strength but with no channel available. Although the handover scheme based on bandwidth has the best performance on new calls blocked rates, its handover calls dropped rates are higher than the other three schemes, as shown in Figure 3. Mobile station always selects the base station with more channels, but the signal strength of the selected base station may not be very strong. In this manner, it may be dropped due to signal strength below acceptable level. The proposed handover scheme based on fuzzy logic outperforms the others

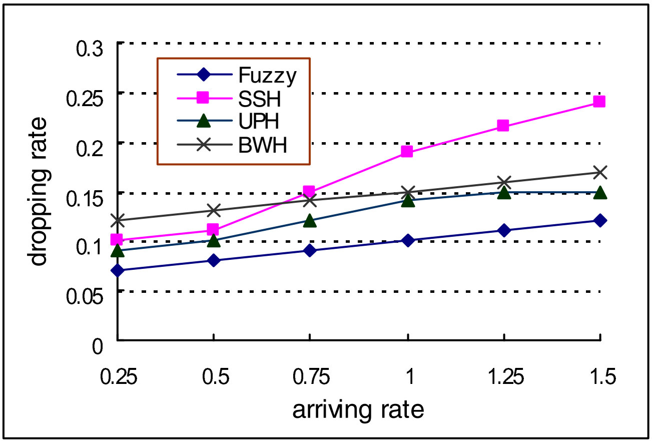


Fig. 3. Dropping rate versus arriving rate

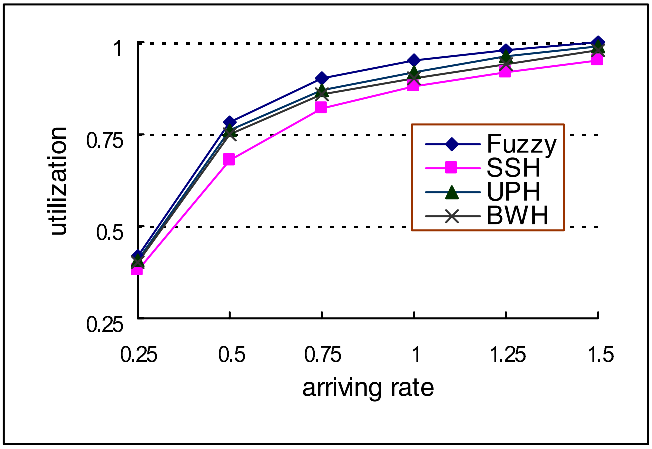


Fig. 4. Utilization versus arriving time

especially in high traffic load because it not only considers the available number of channels but also signal strength, dropping rate, and bandwidth utilization. Figure 4 shows the bandwidth utilization versus mean inter-arrival time. Since signal strength based handover scheme has higher new calls blocking rates and handover calls dropping rates, the bandwidth utilization is naturally lower than the other three schemes.

5 Conclusions

This paper has presented an adaptive fuzzy logic based handover scheme for mobile stations roaming in the WiMAX network. This method takes four criteria into account, including signal strength, number of available channels, bandwidth utilization and handover call dropping rate. The results show that the proposed scheme outperforms the conventional handover schemes based only on the signal strength, user population and bandwidth utilization, respectively.

References

1. Becvar, Z., Zelenka, J.: Implementation of Handover Delay Timer into WiMAX. In: The 6th Conference on Telecommunications, May 2007, pp. 107–114 (2007)
2. Becvar, Z., Zelenka, J.: Handovers in the Mobile WiMAX. In (2008), http://fireworks.intranet.gr/publications/Fireworks_6CTUPB008a.pdf
3. Chan, P.M.L., Hu, Y.F., Sheriff, R.E.: Implementation of Fuzzy Multiple Objective Decision Making Algorithm in a Heterogeneous Wireless Environment. In: Proceeding of IEEE on Wireless Communications and Networking Conference, vol. 1, pp. 332–336 (March 2002)
4. Majiesi, A., Khalaj, B.H.: An Adaptive Fuzzy Logic Based Handover Algorithm for Interworking Between WLANs and Wireless Networks. In: 13th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, vol. 5, pp. 2446–2451 (2002)
5. Chan, P.M.L., Hu, Y.F., Sheriff, R.E.: Mobility Management Incorporating Fuzzy Logic for a Heterogeneous IP Environment. IEEE Communications Magazine 39(12), 42–51 (2001)
6. Pahlavan, K., Krishnamurthy, P., Hatami, A., Ylianttila, M., Makela, J.-P., Pichna, R., Vallstrom, J.: Handover in Hybrid Mobile Data Networks. IEEE Personal Communications, 34–47 (April 2000)
7. Zahariadis, T.B., Vaxevanakis, K.G., Tsantilas, C.P., Zervos, N.A., Nikolaou, N.A.: Global roaming in next-generation networks. IEEE Communications Magazine 40(2), 145–151 (2002)

A Fuzzy Bilevel Model and a PSO-Based Algorithm for Day-Ahead Electricity Market Strategy Making

Guangquan Zhang[1], Guoli Zhang[2], Ya Gao[1], and Jie Lu[1]

[1] Faculty of Engineering & Information Technology,
University of Technology, Sydney, PO Box 123, NSW 2007, Australia
{zhangg,yagao,jielu}@it.uts.edu.au
[2] Department of Mathematics and Physics,
North China Electric Power University, Baoding, 071003, Hebei, P.R. China
zhangguoli@ncepu.edu.cn

Abstract. This paper applies bilevel optimization techniques and fuzzy set theory to model and support bidding strategy making in electricity markets. By analyzing the strategic bidding behavior of generating companies, we build up a fuzzy bilevel optimization model for day-ahead electricity market strategy making. In this model, each generating company chooses the bids to maximize the individual profit. A market operator solves an optimization problem based on the minimization purchase electricity fare to determine the output power for each unit and uniform marginal price. Then, a particle swarm optimization (PSO)-based algorithm is developed for solving problems defined by this model.

Keywords: nonlinear bilevel optimization, fuzzy set, electricity market, strategic bidding, particle swarm algorithm.

1 Introduction

Many decision problems have hierarchical structures, for which multi-level, especially bilevel programming techniques have been developed. In a bilevel decision problem, a decision maker at the upper level is known as the leader, and at the lower level, the follower [1]. Bilevel problems have been studied tremendously and a lot of achievements have been obtained [2][3][4].

Throughout the world, electric power industries are undergoing enormous restructuring processes from nationalized monopolies to competitive market. Because of the significance and particularity of electricity energy to national economics and society, electricity market must be operated under conditions of absolute security and stabilization. The research of electricity market has concerned a lot of researchers, owners and managers from electricity entities and authorities. The competitive mechanism of day-ahead markets is one of very important issues in the electricity market study, which can be described as follows. Each generating company (GC) submits a set of hourly (half-hourly) generation prices and the available capacities for the following day. According to these data and an hourly (half-hourly) load forecast, a market operator (MO) allocates

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 736–744, 2009.

generation output for each unit. A lot of researches have been done on how to strategically bid prices for those GCs, and how to dispatch generation output for MOs for each unit. Literatures [5][6][7] use supply function equilibrium model to describe a day-ahead electricity market to maximize GCs' profits and get Nash equilibrium. Literatures [8][9][10] use game theory to build a strategic bidding model for generating companies, and reach a Nash equilibrium solution. However, these models do not include ramp rate constraints, which are very crucial to guarantee real optimal solutions. In addition, because strategic bidding problems involve two hierarchical optimizations, and are different from a conventional game model, a new Nash equilibrium is needed as a solution. From literatures, only Pang and Fukushima [10] gave a generalized Nash equilibrium concept. Literature [11] used a bilevel optimization method to build a generation output allocation model, but does not consider competitive bidding problem from GCs.

The rest of this paper is organized as follows: Section 2 introduces related definition on fuzzy sets as preliminary of this study. Section 3 analyses and builds a fuzzy bilevel optimization model in day-ahead electricity markets, which includes ramp rate constraints. To solve problems defined by this model, Section 4 provides a solution algorithm. Finally, conclusions and further study are highlighted in Section 5.

2 Preliminary

Definition 2.1[12] Let $\tilde{a},\tilde{b} \in F(R)$ are two fuzzy numbers, the ranking relationship between $\tilde{a}$ and $\tilde{b}$ is defined as follows:

$$\tilde{a} \le \tilde{b} \text{ if } m(\tilde{a}) < m(\tilde{b}) \text{ or } m(\tilde{a}) = m(\tilde{b}) \text{and } \sigma(\tilde{a}) \ge \sigma(\tilde{b})$$

where the mean $m(\tilde{a})$ and the standard deviation $\sigma(\tilde{a})$ are defined as

$$m(\tilde{a}) = \frac{\int_{s(\tilde{a})} x\tilde{a}(x)dx}{\int_{s(\tilde{a})} \tilde{a}(x)dx}$$

$$\sigma(\tilde{a}) = \left(\frac{\int_{s(\tilde{a})} x^2\tilde{a}(x)dx}{\int_{s(\tilde{a})} \tilde{a}(x)dx} - (m(\tilde{a}))^2 \right)^{1/2}$$

where $s(\tilde{a}) = \{x \mid \tilde{a}(x) > 0\}$ is the support set of fuzzy number $\tilde{a}$.

For triangular fuzzy number $\tilde{a} = (l,m,n)$,

$$m(\tilde{a}) = \frac{1}{3}(l+m+n)$$

$$\sigma(\tilde{a}) = \frac{1}{18}(l^2 + m^2 + n^2 - lm - ln - mn)$$

3 Bidding Strategy Analysis in Competitive Electricity Markets

In an auction-based day-ahead electricity market, each GC tries to maximize its own profit by strategic bidding. Specifically, each GC submits a set of hourly generation prices and available capacities for the following day. Based on these data, a MO allocates generation output. This is a typical bilevel decision problem. GCs are leaders, a MO is a follower. In this section, under the analysis of the bidding strategy optimization problem, we build a competitive strategic bidding model for GCs and a generation output dispatch model for a MO in a day-ahead electricity market.

3.1 Generating Companies' Strategic Pricing Model

In the upper level, each GC concerns how to choose a bidding strategy, which includes generation price and available capacity. For a power system, the generation cost function generally adopts a quadratic function of the generation output, i.e. the generation cost function can be represented as

$$C_j(P_j) = a_j P_j^2 + b_j P_j + c_j \tag{1}$$

where P_j is the generation output of generator j, and a_j, b_j, c_j are coefficients of generation cost function of generator j.

The marginal cost of generator j is calculated by

$$\lambda_j = 2a_j P_j + b_j \tag{2}$$

It is a linear function of its generation output P_j. The rule in a goods market may expect each GC to bid according to its own generation cost. Therefore we adopt this linear bid function. Suppose the bidding for j-th unit at time t is

$$R_{tj} = \alpha_{tj} + \beta_{tj} P_{tj} \tag{3}$$

where $t \in T$ is the time interval, T is time interval number, j represents the unit number, P_{tj} is the generation output of unit j at time t, and α_{tj} and β_{tj} are the bidding coefficients of unit j at time t.

According to the Justice Principle of "the same quality, the same network, and the same price", we adopt a UMP as the market clearing price. Once the market is cleared, each unit will be paid according to its generation output and UMP. The payoff of GC$_i$ is

$$F_i = \sum_{t=1}^{T} \left(\sum_{j \in G_i} UMP_t P_{tj} - \sum_{j \in G_i} (a_j P_{tj}^2 + b_{tj} P_{tj} + c_{tj}) \right) \tag{4}$$

where G_i is the suffix set of the units belonged to GC$_i$. Each GC wishes to maximize its own profit F_i. In fact, F_i is the function of P_{tj} and UMP_t, and UMP_t is the function of all units' bidding α_{tj}, β_{tj} and output power P_{tj}, which will impose impact to each other. Therefore, we can establish a strategic pricing model of these GC competitive bidding strategic as follows

$$\max_{\alpha_{tj}, \beta_{tj}, j \in G_i} F_i = F_i(\alpha_{t1}, \beta_{t1}, \cdots, \alpha_{tN}, \beta_{tN}, P_{t1}, \cdots, P_{tN}) = \sum_{t=1}^{T} \left(UMP_t P_{ti} - \sum_{j \in G_i} (a_j P_{tj}^2 + b_j P_{tj} + c_j) \right)$$

$$i = 1, 2, \cdots, L \tag{5}$$

where L is GC number, $P_{ti} = \sum_{j \in G_i} P_{tj}$, $t = 1, 2, \cdots, T$.

The profit calculating for each GC will consider both P_{tj} and UMP_t, which can be computed by MO according to the market clearing model.

In the model above, the cost coefficients are generally obtained from experiment. However, there indeed exist many facts which affect the value of cost coefficients, such as experiment errors, different operation situations, quality of coal and the ageing of facilities. Therefore it would be more reasonable to describe the cost coefficients as fuzzy numbers. Based on this consideration, a more authentic model of the real game model with uncertain cost coefficient for GC competitive strategic bidding is established as follows:

$$\max_{\alpha_{tj},\beta_{tj},j\in G_i} F_i = F_i(\alpha_{t1},\beta_{t1},\cdots,\alpha_{tN},\beta_{tN},P_{t1},\cdots,P_{tN}) = \sum_{t=1}^{T}(UMP_t P_{ti} - \sum_{j\in G_i}(\tilde{a}_j P_{tj}^2 + \tilde{b}_j P_{tj} + \tilde{c}_j)) \tag{6}$$

where L is GC number, $P_{ti} = \sum_{j\in G_i} P_{tj}$, $t = 1,2,\cdots,T$, $\tilde{a}_j,\tilde{b}_j,\tilde{c}_j$ are fuzzy cost coefficients.

3.2 A Market Operator's Generation Output Dispatch Model

A MO actually represents the consumer electricity purchase from GCs, under the conditions of security and stabilization. The objective of a MO is to minimize the total purchase fare while encouraging GCs to bid price as low as possible. It is reasonable that the lower the price, the more the output. Thus, the function value of a MO's objective will be calculated according to the bidding price. Most previous bidding strategic models do not include ramp rate constraints, without which, the solution for generating dispatch may not be a truly optimal one. However, if a model includes ramp rate as constraints, the number of decision variables will increase dramatically, which imposes stronger request for a more powerful solution algorithm. Based on the analysis above, we build a MO's generation output dispatch model as follows:

$$\begin{cases} \min\limits_{P_{tj}} f = f(\alpha_{t1},\beta_{t1},\cdots,\alpha_{tN},\beta_{tN},P_{t1},\cdots,P_{tN}) = \sum\limits_{t=1}^{T}\sum\limits_{j=1}^{N} R_{tj} P_{tj} \\ \sum\limits_{j=1}^{N} P_{tj} = P_{tD} \\ P_{j\min} \le P_{tj} \le P_{j\max} \\ -D_j \le P_{tj} - P_{t-1,j} \le U_j, t = 1,2,\cdots,T \end{cases} \tag{7}$$

where $t \in T$ is the time interval, T is time interval number, *j* represents the unit number, P_{tj} is the generation output of unit j at time t, and α_{tj} and β_{tj} are the bidding coefficients of unit j at time t, P_{tD} is the load demand at time t, $P_{j\min}$ is the minimum output power of the *j*-th unit, $P_{j\max}$ is the maximum output power of the *j*-th unit, D_j is the maximum downwards ramp rate of the *j*-th unit, and U_j is the maximum upwards ramp rate of the *j*-th unit.

After receiving all GCs' bid data, MO determines the output power of each unit and UMP_t for all time slot t. UMP_t can be calculated according to following steps:

Step1: calculate output power of each unit j for all time slot t using model (7);
Step2: compute bidding R_{tj} corresponding to the generation output P_{tj};
Step3: account $UMP_t = \max_{j=1}^{N} R_{tj}$.

3.3 A Fuzzy Bilevel Optimization Model for Strategic Biddings in Electricity Markets

From the analysis in above, we know that in an auction-based day-ahead electricity market, each GC tries to maximize its own profit by strategic bidding, and each MO tries to minimize its total electricity purchase fare. The decision from either of them will influence the other. This is a typical bilevel decision problem, which has multi-leaders and only one follower, with GCs as leaders and a MO as a follower.

By combining the strategic pricing model defined in (6) with the generation output dispatch model defined in (7), we establish a fuzzy bilevel optimization model for competitive strategic bidding-generation output dispatch in an auction-based day-ahead electricity market as follows:

$$\left\{\begin{array}{l}
\max\limits_{\alpha_{tj},\beta_{tj},j\in G_i} F_i = F_i(\alpha_{t1},\beta_{t1},\cdots,\alpha_{tN},\beta_{tN},P_{t1},\cdots,P_{tN}) = \sum\limits_{t=1}^{T}(UMP_tP_{ti} - \sum\limits_{j\in G_i}(\tilde{a}_jP_{tj}^2 + \tilde{b}_jP_{tj} + \tilde{c}_j)) \\
\alpha_{t\min} \le \alpha_{tj} \le \alpha_{t\max}, \beta_{t\min} \le \beta_{tj} \le \beta_{t\max}, t = 1,2,\cdots,T; j = 1,2,\cdots,N; i = 1,2,\cdots,L \\
\qquad \min\limits_{P_{tj}} f = f(\alpha_{t1},\beta_{t1},\cdots,\alpha_{tN},\beta_{tN},P_{t1},\cdots,P_{tN}) = \sum\limits_{t=1}^{T}\sum\limits_{j=1}^{N} R_{tj}P_{tj} \\
\qquad \sum\limits_{j=1}^{N} P_{tj} = P_{tD} \\
\qquad P_{j\min} \le P_{tj} \le P_{j\max} \\
\qquad \text{-}D_j \le P_{tj} - P_{t-1,j} \le U_j, t = 1,2,\cdots,T
\end{array}\right. \tag{8}$$

where α_{tj} and β_{tj} are the bidding coefficients of unit j at time t , $\tilde{a}_j,\tilde{b}_j,\tilde{c}_j$ are fuzzy cost coefficients, L is the number of generating companies, $P_{ti} = \sum_{j\in G_i} P_{tj}$, $P_{j\min}$ is the minimum output power of the j-th unit, $P_{j\max}$ is the maximum output power of the j-th unit, D_j is the maximum downwards ramp rate of the j-th unit, and U_j is the maximum upwards ramp rate of the j-th unit.

This is a bilevel problem with fuzzy parameters. In order to get a solution, we will develop a solution algorithm in next section.

4 A Particle Swarm Optimization Based Algorithm

In this section, we use the strategy adopted in PSO method [13] to develop a PSO-based algorithm to reach a solution for the problem defined by (8) .

Fig.1 outlines the main structure of this algorithm. We first sample the leaders-controlled variables to get some candidate choices for leaders. Then, we deal with fuzzy coefficients by the fuzzy ranking method defined in Definition 2.1. For every leader's choice we use PSO method together with Stretching technology [13] to get the follower's response. Thus a pool of candidate solutions for both the leaders and the follower is formed. By pushing every solution pair moving towards current best ones, the whole solution pool is updated. Once a solution is reached for the leaders, we use Stretching technology to escape local optimization. We repeat this procedure by a pre-defined count and reach a final solution.

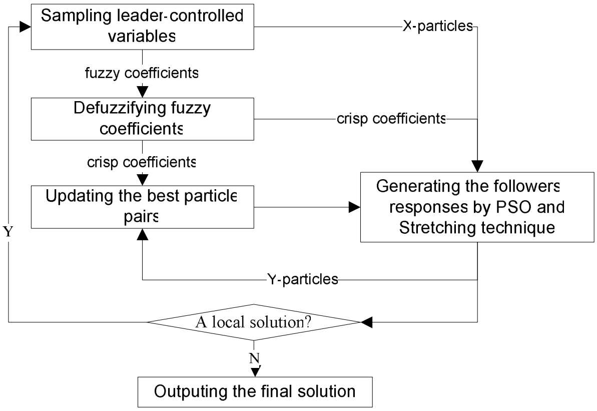

Sampling leader-controlled variables
X-particles
fuzzy coefficients
Defuzzifying fuzzy coefficients
crisp coefficients
crisp coefficients
Updating the best particle pairs
Generating the followers responses by PSO and Stretching technique
Y
Y-particles
A local solution?
N
Outputing the final solution

Fig. 1. The outline of the PSO based algorithm

The detailed PSO-based algorithm has two parts, Algorithm 1, which is to generate the response from a follower, and Algorithm 2, which is to generate optimal strategies for all leaders. These two algorithms are specified as below.

Algorithm 1: Generate the response from a follower

Step 1: Input the values of x_{ij} from the L leaders;

Step 2: Sample N_f candidates y_i and the corresponding velocities v_{y_i}, $i = 1,\ldots,N_f$;

Step 3: Initiate the follower's loop counter k_f =0;

Step 4: Record the best particles p_{y_i} and y^* from p_{y_i}, $i = 1,\ldots,N_f$;

Step 5: Update velocities and positions using

$$v_{y_i}^{k+1} = w_f v_{y_i}^{K} + c_f r_{1l}^K (p_{y_i} - y_i^K) + c_f r_{2l}^K (y^{*K} - y_i^K)$$

$$y_{\mathrm{i}}^{K+1} = y_i^K + v_{y_i}^{k+1}$$

Step 6: $k_f = k_f + 1$;

Step 7: If $k_f \geq MaxK_f$ or the solution changes for several consecutive generations are small enough, then we use Stretching technology to obtain the global solution and go to Step 8. Otherwise go to Step 5;

Step 8: Output y^* as the response from the follower.

Algorithm 2: Generate optimal strategies for leaders

```
Step 1: Sample $N_l$ particles of $x_{ij}$, and the corresponding velocities $v_{x_{ij}}$;
Step 2: Initiate the leaders' loop counter $k_l = 0$;
Step 3: For the k-th particle, $k = 1,\ldots, N_l$, calculate the optimal response $r_i(x_{-i})$ by Formula (5), $i = 1,\ldots,L$;
    Step 3.1 : Sample $N_l$ particles $x_{ij}$ within the constraints of $x_{ij}$;
    Step 3.2 : By calling Algorithm 1, we calculate the rational response from the follower;
    Step 3.3 : Using PSO technique, we obtain $r_i(x_{-i})$; $i = 1,\ldots,L$;
Step 4: By using the fuzzy ranking method defined in Definition 2.1, calculate the function value of every particle by Formula (6);
Step 5: Record $p_{x_{ij}}$, $x_{ij}^*$, $j = 1,\ldots, N_l$ for each $x_{ij}$, $j = 1,\ldots, N_l$
Step 6: Update velocities and positions using
            $v_{x_{ij}}^{k+1} = w_l v_{x_{ij}}^{K} + c_l r_{1l}^K (p_{x_{ij}} - x_{ij}^K) + c_l r_{2l}^K (x_{ij}^{*K} - x_{ij}^K)$
            $x_{\mathrm{ij}}^{K+1} = x_{ij}^K + v_{x_{ij}}^{k+1}$
Step 7: $k_l = k_l + 1$;
Step 8: If $\sum_{i=1}^{L} | x_i - r_i(x_{-i}) | \le \varepsilon$ or $k_l \ge MaxK_l$, then we use Stretching technology to current leaders' solutions to obtain the global solution.
```

Notations used in the algorithms are detailed in Table 1.

Table 1. Explanation of some notations used in the PSO-based algorithm

N_l	the number of candidate solutions (particles) for leaders
N_f	the number of candidate solutions (particles) for the follower
x_{ij}	the *j*-th candidate solutions for the controlling variables from *i*-th leader
$p_{x_{ij}}$	the best previously visited position of x_{ij}
x_{ij}^*	current best one for particle x_{ij}
$v_{x_{ij}}$	the velocity of x_{ij}
k_l	current iteration number for the upper-level problem
y_i	*i*-th candidate solution for the controlling variables from the follower
p_{y_i}	the best previously visited position of y_i
y^*	current best one for particle *y*
v_{y_i}	the velocity of y_i
k_f	current iteration number for the lower-level problem
$MaxK_l$	the predefined max iteration number for k_l

Table 1. (*continued*)

$MaxK_f$	the predefined max iteration number for k_f
w_l, w_f	inertia weights for leaders and follower respectively (coefficients for PSO)
c_l, c_f	acceleration constants for leaders and follower respectively (coefficients for PSO)

5 Conclusions

Based on the analysis for strategic bidding behaviors in a day-ahead electricity market, this paper builds up a fuzzy bilevel optimization model for it. The PSO-based algorithm given in this paper is to obtain solutions for competitive strategic bidding problems, and to provide GCs competitive strategic biddings within network security constraints.

Further theory research and experiment analysis will be carried to study the bidding strategy in day-ahead electricity markets through collecting more data from real world.

Acknowledgment. The work presented in this paper was supported by Australian Research Council (ARC) under discovery project DP0557154.

References

[1] Zhang, G., Lu, J., Gao, Y.: An algorithm for fuzzy multi-objective multi-follower partial cooperative bilevel programming. International Journal of Intelligent & Fuzzy Systems 19, 303–319 (2008)

[2] Zhang, G., Lu, J., Gao, Y.: Fuzzy bilevel programming: Multi-objective and multi-follower with shared variables. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 16, 105–133 (2008)

[3] Sakawa, M., Nishizaki, I.: Interactive fuzzy programming for two-level nonconvex programming problems with fuzzy parameters through genetic algorithms. Fuzzy Sets and Systems 127, 185–197 (2002)

[4] Lu, J., Shi, C., Zhang, G.: An extended branch-and-bound algorithm for bilevel multi-follower decision making in a referential-uncooperative situation. International Journal of Information Technology and Decision Making 6, 371–388 (2006)

[5] Rudkevich, A.: Supply Function Equilibrium: Theory and Applications. In: Proceedings of the 36th Annual Hawaii International Conference on System Sciences, January 6-9 (2003)

[6] Tao, L.: Mohammad Shahidehpour, Strategic bidding of transmission-constrained GENCOs with incomplete information. IEEE Transactions on Power Systems 20, 437–447 (2005)

[7] Haghighat, H., Seifi, H., Ashkan, R.K.: Gaming Analysis in Joint Energy and Spinning Reserve Markets. IEEE Transactions on Power Systems 22, 2074–2085 (2007)

[8] Wen, F., Kumar, A.: Optimal bidding strategies and modeling of imperfect information among competitive generators. IEEE Transactions on Power Systems 16, 15–21 (2001)

[9] James, D.W., Thomas, J.O., James, D.W., Thomas, J.O.: An Individual Welfare Maximization Algorithm for Electricity Markets. IEEE Transactions on Power Systems 17, 590–596 (2002)

[10] Pang, J., Fukushima, M.: Quasi-variational inequalities, generalized Nash equilibria, and multi-leader-follower games. Computational Management Science 2, 21–56 (2005)

[11] Bjøndal, M., Jørnsten, K.: The Deregulated Electricity Market Viewed as a Bilevel Programming Problem. Journal of Global Optimization 33, 465–475 (2005)

[12] Lee, E.S., Li, R.L.: Comparison of fuzzy numbers based on the probability measure of fuzzy events. Comput. Math. Appl. 15, 887–896 (1988)

[13] Parsopoulos, K.E., Vrahatis, M.N.: Recent approaches to global optimization problems through particle swarm optimization. Natural Computing 1, 235–306 (2002)

Correspondence between Incomplete Fuzzy Preference Relation and Its Priority Vector

Pei-Di Shen[1], Wen-Li Chyr[2], Hsuan-Shih Lee[3], and Kuang Lin[3]

[1] Graduate School of Education
[2] Department of Information Management
Ming Chung University
Taipei 111, Taiwan
[3] Department of Shipping and Transportation Management
National Taiwan Ocean University
Keelung 202, Taiwan

Abstract. Fuzzy preference relations are frequently adopted by decision makers to express their preference tendency toward alternatives. Due to the lack of expertise of knowledge, decision makers may not be able to specify complete preference relation. To deal with incomplete fuzzy preference relations, Xu [26] proposed prioritization methods for incomplete fuzzy preference relations where he postulated a correspondence between priority vector and additive consistent incomplete fuzzy preference relation. In this paper, we are going to prove the correspondence does not always hold.

Keywords: consistent incomplete fuzzy preference relation, incomplete fuzzy preference relation, priority vector.

1 Introduction

Much research has been devoted to develop new methods to facilitate the decision process for decision makers. Introducing of the fuzzy theory into decision model is one of these advancements. Different models have been proposed for decision-making problems under fuzzy environment [15-18]. Decision-making process usually consists of multiple individuals interacting to reach a decision. Different experts may express their evaluations by means of different preference representation formats and as a result different approaches to integrating different preference representation formats have been proposed [1,2,8,10,29,30]. In these research papers, many reasons are provided for fuzzy preference relations to be chosen as the base element of that integration.

One important issue of fuzzy preference relation is that of "consistency" [3,4,11]. Many properties have been suggested to model transitivity of fuzzy preference relations and some of these suggested properties are as follows:

(1) Triangle condition [11,19]
(2) Weak transitivity [11,23]
(3) Max-min transitivity [11,28]

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 745–751, 2009.

(4) Max-max transitivity [7,11,28]
(5) Restricted max-min transitivity [11,23]
(6) Restricted max-max transitivity [11,23]
(7) Additive transitivity [11,19,23]
(8) Multiplicative transitivity [11,22,23,25,27]

Amongst these properties two of them attract more attentions in recent research [11,25,26], which are additive transitivity and multiplicative transitivity.

One of research focuses of fuzzy preference relations is to prioritize alternatives based on the fuzzy preference relation. Many methods have been proposed to draw priorities from a multiplicative preference relation, such as the eigenvector method [22], the least square method [12], gradient eigenvector method [5], logarithmic least square method [6] and generalized chi square method [24], etc. When using fuzzy preference relations, some priority methods have been given using what have been called choice functions or degrees [1,9,13,14,20,21].

Another important research issue of fuzzy preference relations is how to draw consistent preferences when the fuzzy preference relations are incomplete. Xu [26] has proposed two goal programming models, based on additive consistent incomplete fuzzy preference relation and multiplicative consistent incomplete fuzzy preference relation, for obtaining the priority vector of incomplete fuzzy preference relations. In Xu's method based on additive consistency, he postulated a correspondence between priority vector and additive consistent incomplete fuzzy preference relation. We are going to provide formal proof that the correspondence does not always hold..

2 Preliminaries

For simplicity, we let $N = \{1,2,\ldots,n\}$.

Definition 2.1. Let $R = (r_{ij})_{n\times n}$ be a preference relation, then R is called a fuzzy preference relation [2,13,23], if

$$r_{ij} \in [0,1], \qquad r_{ij} + r_{ji} = 1, \qquad r_{ii} = 0.5 \qquad \text{for all } i, j \in N .$$

Definition 2.2. Let $R = (r_{ij})_{n\times n}$ be a fuzzy preference relation, then R is called an additive consistent fuzzy preference relation, if the following additive transitivity (given by Tanino [23]) is satisfied:

$$r_{ij} = r_{ik} - r_{jk} + 0.5, \qquad \text{for all } i, j, k \in N$$

Definition 2.3. Let $R = (r_{ij})_{n\times n}$ be a fuzzy preference relation, then R is called a multiplicative consistent fuzzy preference relation, if the following multiplicative transitivity (given by Tanion [23]) is satisfied:

$$r_{ik} r_{kj} r_{ji} = r_{ki} r_{ij} r_{jk} \qquad \text{for all } i, j, k \in N$$

Xu extended the concepts in previous section to the situations where the preference information given by the DM (decision maker) is incomplete.

Definition 2.4. [26] Let $R=(r_{ij})_{n\times n}$ be a preference relation, then R is called an incomplete fuzzy preference relation, if some of its elements cannot be given by the DM, which we denote by the unknown number x, and the others can be provided by the DM, which satisfy

$$r_{ij}\in[0,1],\qquad r_{ij}+r_{ji}=1,\qquad r_{ii}=0.5.$$

Definition 2.5. [26] Let $R=(r_{ij})_{n\times n}$ be an incomplete fuzzy preference relation, then R is called an additive consistent incomplete fuzzy preference relation, if all the known elements of R satisfy the additive transitivity

$$r_{ij}=r_{ik}-r_{jk}+0.5.$$

Definition 2.6. [26] Let $R=(r_{ij})_{n\times n}$ be an incomplete fuzzy preference relation, then R is called a multiplicative consistent incomplete fuzzy preference relation, if all the known elements satisfy the multiplicative transitivity $r_{ik}r_{kj}r_{ji}=r_{ki}r_{ij}r_{jk}$.

For the convenience of computation, Xu constructed an indication matrix $\Delta=(\delta_{ij})_{n\times n}$ of the incomplete fuzzy preference relation $R=(r_{ij})_{n\times n}$, where

$$\delta_{ij}=\begin{cases}0, & r_{ij}=x\\ 1, & r_{ij}\neq x.\end{cases}$$

Xu [26] developed two goal programming models, based on additive consistent incomplete fuzzy preference relation and multiplicative consistent incomplete fuzzy preference relation respectively, for obtaining the priority vector of incomplete fuzzy preference relation.

Let $w=(w_1,w_2,\ldots,w_n)^T$ be the priority vector of the incomplete fuzzy preference relation $R=(r_{ij})_{n\times n}$, where $w_i\geq 0$, $i\in N$, $\sum_{i=1}^{n} w_i=1$.

Xu postulated that

(1) If $R=(r_{ij})_{n\times n}$ is an additive consistent incomplete fuzzy preference relation, then such a preference relation is given by

$$\delta_{ij}r_{ij}=\delta_{ij}[0.5(w_i-w_j+1)],\qquad i,j\in N. \tag{1}$$

(2) If $R=(r_{ij})_{n\times n}$ is a multiplicative consistent incomplete fuzzy preference relation, then such a preference relation is given by

$$\delta_{ij}r_{ij}=\delta_{ij}\frac{w_i}{w_i+w_j}\qquad i,j\in N. \tag{2}$$

Based on (1), Xu constructed the following multi-objective programming model to obtain the priority vector:

$$\begin{aligned}&\text{(MOP1)}\ \min\ \varepsilon_{ij}=\delta_{ij}\,|\,r_{ij}-0.5(w_i-w_j+1)\,|, \qquad i,j\in N\\ &s.t.\ \ w_i\ge 0,\quad i\in N,\ \sum_{i=1}^{n}w_i=1\end{aligned} \tag{3}$$

Based on (2), Xu formulated another multi-objective programming model to derive the priority vector:

$$\begin{aligned}&\text{(MOP2)}\ \min\ \varepsilon_{ij}=\delta_{ij}\,|\,r_{ij}w_j-r_{ij}w_i\,|, \qquad i,j\in N\\ &s.t.\ \ w_i\ge 0,\quad i\in N,\ \sum_{i=1}^{n}w_i=1\end{aligned} \tag{4}$$

3 Relations between Fuzzy Preference Relation and Priority Vector

Xu [26] postulated that an additive consistent incomplete fuzzy preference relation $R=(r_{ij})_{n\times n}$ satisfies (1). However, (1) does not hold for all additive consistent fuzzy preference relation, which is proven in the following theorem.

Theorem 3.1 For any $n\ge 3$, there exits incomplete fuzzy preference relation $R=(r_{ij})_{n\times n}$ such that R is an additive consistent incomplete fuzzy preference relation but violates (1).

<Proof> Let $R=(r_{ij})_{n\times n}$ be an incomplete fuzzy preference relation where

$$r_{ij}=\begin{cases}x & i=2 \text{ and } j>2,\\ (\dfrac{i-j}{n-1}+1)/2 & \text{otherwise.}\end{cases}$$

Note that for all known r_{ij} we have $0\le r_{ij}=(\dfrac{i-j}{n-1}+1)/2\le 1$ and $r_{ii}=0.5$. For all known r_{ij} r_{ik} r_{jk} , we have

$$\begin{aligned}&r_{ik}-r_{jk}+0.5=(\frac{i-k}{n-1}+1)/2-(\frac{j-k}{n-1}+1)/2+0.5\\ &=(\frac{i-j}{n-1}+1)/2\\ &=r_{ij}\end{aligned}.$$

Following the definition 2.5, $R=(r_{ij})_{n\times n}$ is an additive consistent incomplete fuzzy preference relation. We find that the priority vector $w=(w_1,w_2,\ldots,w_n)^T$ that satisfies

$$\delta_{ij} r_{ij} = \delta_{ij}[0.5(w_i - w_j + 1)], \qquad i, j \in N,$$

would be $w_i = c + \frac{i-1}{n-1}$, where $1 \le i \le n$ c is a nonnegative number . Since

$$\sum_{i=1}^{n} w_i = nc + \frac{n}{2} > 1,$$

it is impossible to find $w_i \ge 0$ such that $\sum_{i=1}^{n} w_i = 1$. We can conclude that for the additive consistent incomplete fuzzy preference relation R , it is impossible to find the priority vector $w = (w_1, w_2, \dots, w_n)^T$ such that $\delta_{ij} r_{ij} = \delta_{ij}[0.5(w_i - w_j + 1)], \qquad i, j \in N$, where $w_i \ge 0$ and $\sum_{i=1}^{n} w_i = 1$. □

Example 3.1. Consider the following incomplete fuzzy preference relation for four alternatives:

$$R_1 = \begin{bmatrix} 0.5 & 0.25 & 0.125 & 0 \\ 0.75 & 0.5 & x & x \\ 0.875 & x & 0.5 & 0.375 \\ 1 & x & 0.625 & 0.5 \end{bmatrix}.$$

It is very easy to verify that all the known elements of R_1 satisfy the additive transitivity

$$r_{ij} = r_{ik} - r_{jk} + 0.5 .$$

Following the definition 2.5, R_1 is called an additive consistent incomplete fuzzy preference relation. We find that the priority vector $w = (w_1, w_2, w_3, w_4)^T$ that satisfies $\delta_{ij} r_{ij} = \delta_{ij}[0.5(w_i - w_j + 1)], \qquad i, j \in N$ would be $w_1 = c, w_2 = c + 0.5, w_3 = c + 0.75, w_4 = c + 1$, where c is a nonnegative number. Since $w_1 + w_2 + w_3 + w_4 = 2.25 + 4c \ge 0$, it is impossible to find $w_i \ge 0$ such that $\sum_{i=1}^{n} w_i = 1$ and $\delta_{ij} r_{ij} = \delta_{ij}[0.5(w_i - w_j + 1)], \qquad i, j \in N$.

4 Conclusion

We have provided a formal proof that the correspondence in equation (1) is not always valid for additive consistent incomplete fuzzy preference relations. The proof provides insight of the nature of additive consistency, i.e., the difficulty in prioritization assuming the weights are additive. In other words, the prioritization of additive incomplete fuzzy preference relations needs more treatment than other types of incomplete fuzzy preference relations.

Acknowledgement. This research work was partially supported by the National Science Council of the Republic of China under grant No. NSC96-2416-H-019-004-MY2 and grant No. NSC 97-2221-E-019 -022 -.

References

1. Chiclana, F., Herrera, F., Herrera-Viedma, E.: Integrating three representation models in fuzzy multipurpose decision making based on fuzzy preference relations. Fuzzy Sets and Systems 97, 33–48 (1998)
2. Chiclana, F., Herrera, F., Herrera-Viedma, E.: Integrating multiplicative preference relations in a multipurpose decision-making model based on fuzzy preference relations. Fuzzy Sets and Systems 122, 277–291 (2001)
3. Chiclana, F., Herrera, F., Herrera-Viedma, F.E.: Reciprocity and consistency of fuzzy preference relations. In: De Baets, B., Fodor, J. (eds.) Principles of Fuzzy Preference Modelling and Decision Making, pp. 123–142. Academia Press (2003)
4. Chiclana, F., Herrera, F., Herrera-Viedma, F.E.: Rationality of induced ordered weighted operators based on the reliability of the source of information in group decision-making. Kybernetika 40, 121–142 (2004)
5. Cogger, K.O., Yu, P.L.: Eigenweight vectors and least-distance approximation for revealed preference in pairwise weight ratios. Journal of Optimization Theory and Application 46, 483–491 (1985)
6. Crawford, G., Williams, C.: A note on the analysis of subjective judgement matrices. Journal of Mathematical Psychology 29, 387–405 (1985)
7. Dubois, D., Prade, H.: Fuzzy Sets and Systems: Theory and Application. Academic Press, New York (1980)
8. Fan, Z.-P., Xial, S.-H., Hu, G.-F.: An optimization method for integrating tow kinds of preference information in group decision-making. Computers & Industrial Engtineering 46, 329–335 (2004)
9. Herrera, F., Herrera-Viedma, E., Verdegay, J.L.: A sequential selection process in group decision-making with linguistic assessment. Information Sciences 85, 223–239 (1995)
10. Herrera, F., Martinez, L., Sanchez, P.J.: Managing non-homogeneous information in group decision making. European Journal of Operational Research 166, 115–132 (2005)
11. Herrera-Viedma, E., Herrera, F., Chiclana, F., Luque, M.: Some issues on consistency of fuzzy preference relations. European Journal of Operational Research 154, 98–109 (2004)
12. Jensen, R.E.: An alternative scaling method for priorities in hierarchical structures. Journal of Mathematical Psychology 28, 317–332 (1984)
13. Kacprzyk, J.: Group decision making with a fuzzy linguistic majority. Fuzzy Sets and Systems 18, 105–118 (1986)

14. Kacprzyk, J., Roubens, M.: Non-Conventional Preference Relations in Decision-Making. Springer, Berlin (1988)
15. Lee, H.-S.: Optimal consensus of fuzzy opinions under group decision making environment. Fuzzy Sets and Systems 132(3), 303–315 (2002)
16. Lee, H.-S.: On fuzzy preference relation in group decision making. International Journal of Computer Mathematics 82(2), 133–140 (2005)
17. Lee, H.-S.: A Fuzzy Method for Measuring Efficiency under Fuzzy Environment. In: Khosla, R., Howlett, R.J., Jain, L.C. (eds.) KES 2005. LNCS, vol. 3682, pp. 343–349. Springer, Heidelberg (2005)
18. Lee, H.-S.: A Fuzzy Multi-Criteria Decision Making Model for the Selection of the Distribution Center. In: Wang, L., Chen, K., S. Ong, Y. (eds.) ICNC 2005, vol. 3612, pp. 1290–1299. Springer, Heidelberg (2005)
19. Luce, R.D., Suppes, P.: Preference utility and subject probability. In: Luce, R.D., et al. (eds.) Handbook of Mathematical Psychology, vol. III, pp. 249–410. Wiley, New York (1965)
20. Orlovvsky, S.A.: Decision making with a fuzzy preference relation. Fuzzy Sets and Systems 1, 155–167 (1978)
21. Roubens, M.: Some properties of choice functions based on valued binary relations. European Journal of Operational Research 40, 309–321 (1989)
22. Saaty, T.L.: The Analytic Hierarchy Process. McGraw-Hill, New York (1980)
23. Tanino, T.: Fuzzy preference orderings in group decision-making. Fuzzy Sets and Systems 12, 117–131 (1984)
24. Xu, Z.S.: Generalized chi square method for the estimation of weights. Journal of Optimization Theory and Applications 107, 183–192 (2002)
25. Xu, Z.S.: Two methods for ranking alternatives in group decision-making with different preference information. Information: An International Journal 6, 389–394 (2003)
26. Xu, Z.S.: Goal programming models for obtaining the priority vector of incomplete fuzzy preference relation. International Journal of Approximate Reasoning 36, 261–270 (2004)
27. Xu, Z.S., Da, Q.L.: An approach to improving consistency of fuzzy preference matrix. Fuzzy Optimization and Decision Making 2, 3–12 (2003)
28. Zimmermann, H.J.: Fuzzy Set Theory and Its Applications. Kluwer, Dordrecht (1991)
29. Zhang, Q., Chen, J.C.H., He, Y.-Q., Ma, J., Zhou, D.-N.: Multiple attribute decision making: approach integrating subjective and objective information. International Journal of Manufacturing Technology and Management 5(4), 338–361 (2003)
30. Zhang, Q., Chen, J.C.H., Chong, P.P.: Decision consolidation: criteria weight determination using multiple preference formats. Decision Support Systems 38, 247–258 (2004)

Nature Inspired Design of Autonomous Driving Agent — Realtime Localization, Mapping and Avoidance of Obstacle Based on Motion Parallax

Ivan Tanev and Katsunori Shimohara

Department of Information Systems Design, Doshisha University,
1-3 Miyakodani, Tatara, Kyotanabe 610-0321 Japan
{itanev,kshimoha}@mail.doshisha.ac.jp

Abstract. We present an approach for nature-inspired design of the driving style of an agent, remotely operating a scale model of a car with obstacle avoidance capabilities. The agent perceives the position of the car from an overhead video camera and conveys its actions to the car via standard radio control transmitter. In order to cope with the video feed latency we propose an anticipatory modeling in which the agent considers its current actions based on the anticipated intrinsic (rather than currently available, outdated) state of the car and its surrounding. Moreover, in a real-time the agent is able (i) to detect a static obstacle with a priori unknown coordinates using onboard video camera, (ii) to map the global position of the obstacle in a nature-inspired way by observing the dynamics of the change of visual angle (i.e., the motion parallax) of the obstacle in several consecutive video frames, and, (iii) in the vicinity of the latter, to employ a potential field-based obstacle avoidance maneuver. Presented work could be seen as a step towards the automated design of the control software of remotely operated vehicles capable to find a safe solution in changeable and uncertain environments.

1 Introduction

The success of the computer playing sport games has long served as touchstone of the progress in the filed of artificial intelligence (AI) [3][4]. The expanding scope of applicability of AI, when the latter is employed to control the individual characters (agents) which are able to "learn" the environment and to adopt an adaptive optimal (rather than a priori preprogrammed) playing tactics and strategy include soccer [6], F1 racing [10], Motocross racing, etc. Focusing in the domain of car racing, in this work we consider the problem of designing a driving agent, able to remotely control a scale model of a car, which runs in a safest possible way around a predefined circuit with an obstacle featuring a priori unknown position.

The *objective* of our work is a nature inspired design of the functionality of driving agent, able to remotely operate a scale model of racing car (hereafter referred to as "car"). The agent should be able to control the car in (i) a consistent way and (ii) to avoid small, static obstacle situated in the driving line at a priory unknown position. An agent with such capabilities would open up an opportunity for building a

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 752–760, 2009.

framework of a novel racing games in which the human competes against a computer with both of them remotely operating scale models, rather than simulated cars. The proposed approach could also be applied for the design of the control software of remotely operated robust vehicles capable to find a safe solution to various tasks in different environmental situations and conditions.

Achieving the objective implies that the following tasks should be addressed: (i) developing an approach that allows the agent to adequately control the scale model of the car addressing the challenge of controlling a fast moving artifact via closed control loop with a finite feedback latency; (ii) formalizing the driving style and defining the key parameters that describe it, and (iii) developing an algorithm paradigm for automated definition of the fastest driving style by setting its key parameters to their optimal values.

Achieving the objective implies that the following tasks should be addressed:

(i) Developing an approach allowing the agent to adequately control the scale model of the car addressing the challenge of dealing with the control feedback latency,

(ii) Formalizing the notion of driving style with obstacle avoidance capabilities and defining the key parameters which describe it, and

(iii) Localization of the obstacle, mapping its position in the global scene, and performing the obstacle avoidance maneuver in a real time.

The related research done by Wloch and Bentley [10] demonstrates the feasibility of applying genetic algorithms for automated optimization of the setup of the simulated racing car. Togelius and Lucas [8] used scale models of cars in their research to demonstrate the ability of the artificial evolution to develop optimal neuro-controllers with various architectures. However, the effects of the inherent latencies in the video feedback on either the precision or the speed of the car was beyond the scope of their work. In our previous work [9], we discuss an evolutionary approach to optimize the controller of a scale model of a racing car capable to avoid a small obstacle situated in the optimized (via genetic algorithms) driving line. However, the position of the obstacle in the global scene is considered was a priory known, and consequently, the real time detection and mapping was not investigated in the considered research.

The remaining of the article is organized as follows. Section 2 introduces the hardware configuration used in our work. It also elaborates on the formalization of the driving style and the adequacy of the control of the car with latency feedback. Section 3 discusses the nature inspired approach based on motion parallax, for real time detection, mapping and avoidance of a small obstacle. This section also presents the experimental results. Section 4 draws a conclusion.

2 System Configuration

2.1 The Car

In our experiments we choose the 1:20, off-the-shelf scale model of a car., which features a digital-proportional radio remote control (RC) with functionality including "forward", "reverse" (both with controllable velocity) and "neutral" throttle control commands and "left", "right" (both with controllable steering angle) and "straight" steering controls. The car has the following two favorable features: (i) a wide steering angularity, and (ii) a differential drive. The former feature implies a reduced turning angle, and consequently, high maneuverability of the car. The latter feature - differential rear wheels drive (similar to the real cars') implies that that the torque of the motor is split and delivered to the rear wheels in a way that allows them to rotate at different angular speeds if necessary, e.g., under cornering. Therefore, the car turns without a rear wheels spin, which results in a smooth entrance into the turns and a good traction at their exits.

2.2 Perceptions and Actions of the Driving Agent

The driving agent, which controls the car, perceives the current state (location, orientation, and speed) of the car from an overhead video camera. The camera features a high definition CCD sensor and lenses with wide field of view (66 degrees), which allows to cover a sufficiently wide area of about 3500mm x 1800mm from an altitude of about 2500mm. In our experiments camera operates at 640x480 pixels mode, with a video sampling interval of about 33ms. The camera is connected to the personal computer (PC) through a video capture board.

The state of the car is perceived by the onboard camera as illustrated in Figure 2. The car actively emits an infrared (IR) light which is used by overhead camera to track the current position of the car. Such active IR tracking of the car contributes to both a better reliability and precision compared to the previously implemented passive tracking in the visible light spectrum. The latter proved to be prone to significant error due to changeable light condition and variable foreshortening. A snapshot of view of overhead camera, with a car tracked by its IR emission, is shown in Figure 3.

The agent's actions (a series of steering and throttle commands) are conveyed to the car via standard radio control transmitter operating in 27MHz band. The two-channel digital proportional steering and throttle controls are modeled as a computer-generated 8 bit digital signal (4 bits per channel), converted by analog-to-digital converter (ADC) to analog values directly fed to the RC unit. These analog values model the electrical potential normally obtained from the human-operated control potentiometers of the RC unit. The ADC is mounted on a small interface board, connected to the parallel (LPT) port of the PC.

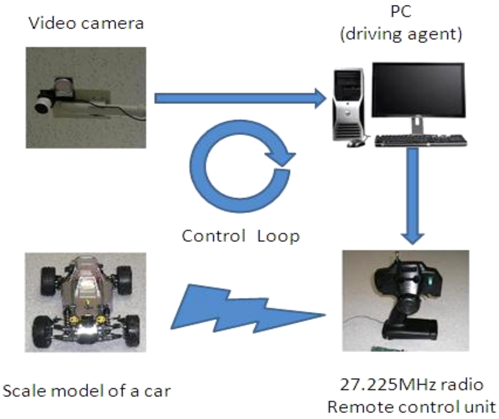


Fig. 1. System configuration

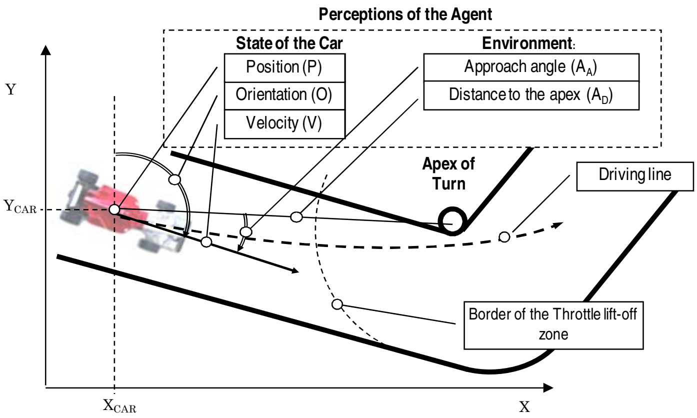


Fig. 2. State of the car and environment perceived by the driving agent

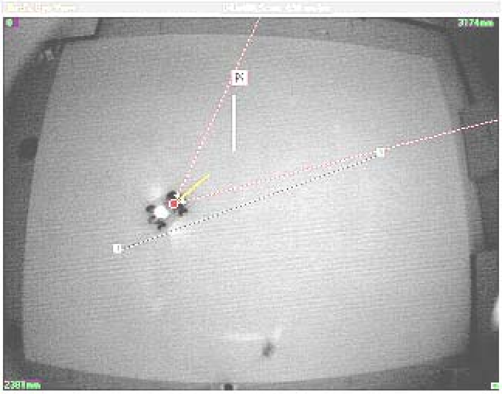


Fig. 3. Snapshot of the overhead camera view of the car, tracked by IR light

2.3 Parameters of the Driving Style

We consider the driving style as the driving line, which the car follows before, around, and after the turns in the circuits combined with the speed, at which the car travels along this line. Our choice of driving styles' parameters is based on the view, shared among the drivers from various teams in different formulas of high performance and competitive driving, that (i) the track can be seen as a set of consequent turns they need to optimize divided by simple straights and that (ii) the turns and both the preceding and following straights should be treated as a single whole [2]. Therefore, we introduce the following key attributes of the driving style: (i) straight-line gear - the gear at which the car approaches the turn, (ii) turning gear, (iii) throttle lift-off zone – the distance from the apex at which the car begins slowing down from the velocity corresponding to the straight line gear to the velocity of the turning gear, (iv) braking velocity - the threshold, above which the car being in the throttle lift-off zone, applies brakes (i.e., reverse throttle command) for slowing down, and (v) approach (homing) angle – the bearing of the apex of the turn. Higher values of the latter parameter yield wider driving lines featuring higher turning radiuses.

2.4 Automated Control of the Car by Driving Agent

Viewing the desired values of driving style parameters as values that the agent needs to maintain, the functionality of the agent includes a continuously perceiving of the state (location, orientation, and speed) of the car and the environment (via onboard camera), computing the error between the values of these parameters and the a priory defined, desired values, and issuing such steering and throttle commands that would minimize this error.

2.5 Anticipatory Modeling

The delays introduced in the feedback control loop (shown in Figure 1) by the latency of the video feed imply that the current actions of the driving agents are based on outdated perceptions, and consequently, outdated knowledge about its own state and the surrounding environment. For the hardware used in our system, the aggregated latency is about 100ms, which results in a maximum error of perceiving the position of the car of about 200mm (scaled error of 4m) when the later runs at its maximum speed of 2000mm/s (scaled speed of 172km/h). The latency also causes an error in perceiving the orientation (bearing) and the speed of the car. The approach is somehow related to the dead reckoning in GPS-based vehicle navigation [1]. As demonstrated in [9], the cumulative effect of these errors renders the tasks of precisely following even simple O-, 8-, and S-shaped routes hardly solvable.

In the proposed approach of incorporating an anticipatory modeling [7], the driving agent considers its current actions based on *anticipated* intrinsic (rather than currently available, outdated) state of the car and surrounding environment. The driving agent *anticipates* the intrinsic state of the car (position, orientation, and speed) from the currently available outdated (by 100ms) state by means of iteratively applying the history of its own most recent actions (i.e., the throttle and steering commands) to the internal model of the car. It further anticipates the perception information related to the surrounding environment, (e.g., the distance and the bearing to the apex of the next turn) from the viewpoint of the anticipated intrinsic position and orientation of the car.

3 Realtime Detection, Mapping and Avoidance of Obstacle

Obstacle avoidance is a key capability of any mobile robot. However, depending on *what* the characteristics of the obstacle are (large or small, static or moving), *whether* the artifact is a priori aware of it or not, and *when* it is introduced to the scene (before the trial or at runtime), the implementation of obstacle avoidance requires an addressing of numerous algorithmic and technological challenges. In this work we consider the simplest case of a *static* obstacle with a priori *unknown* properties (position and size), introduced to the scene *during* the time trial. Such an obstacle, prior to being circumnavigated, should be detected in a realtime and mapped in the global scene. The following subsections are intended to elaborate on this issues.

3.1 Obstacle Detection

In order to detect the obstacle we use a wireless video camera mounted onboard of the car. The camera features a 70 degrees visual field, and perceives the visual angle of the obstacle (if detected) as a simple, one-dimensional estimation of the relative position of the obstacle regarding the car. In order to prevent the scanning of the entire visual field of the camera in the search for obstacle, we employ an anticipatory tracking of the latter, which implies that the agent scans only that part of the entire visual field, which corresponds to the area the obstacle is most likely to be detected. This area in the visual field of the camera is anticipated from the dynamics of the relative movement of the obstacle in the visual field of the onboard camera during the most recent video frames. The scanned area in the visual field of onboard camera is shown by the narrow frame marked by dashed white line in Figure 4.

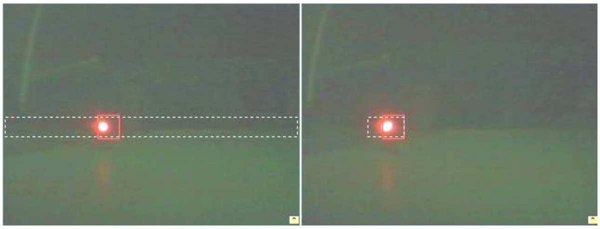


Fig. 4. The scanned area (white dashed frame) in the visual field of onboard camera corresponding to the canonical (left) and anticipatory tracking of the obstacle (right)

3.2 Nature Inspired Mapping of the Obstacle

The mapping of the obstacle is intended to position the obstacle, detected in the visual field of the onboard camera, into the global scene. In order to implement such mapping, we propose a nature inspired mapping based on motion parallax of the obstacle as illustrated in Figure 5. Similar mechanism is employed by most insects in nature for efficient navigation of challenging terrains. As Figure 5 illustrates, the global position (Car_x and Car_y) and orientation (Car_a), integrated with the relative bearing of the obstacle (Cam_a) for two instants of the moving car are sufficient for the determination of the absolute position of the obstacle in the global scene (Obst_x and Obst_y) by means of dynamic triangulation.

As the real data, involved in the computation of the absolute position of the obstacle are inherently noisy, we employ a linearly approximating filtering of these information. Moreover, in order to additionally minimize the error in the determination of the position of the obstacle, the activation of the algorithm for determination of the position of the obstacle is computer only when the angle between the position of the car for the two instants viewed from the obstacle (angle Beta, as shown in Figure 5) is wide enough. The sufficiency of the value of this angle is guaranteed by the allowed time interval between the two instants. This time interval in our approach is set to 20 time steps, which corresponds to 660ms.

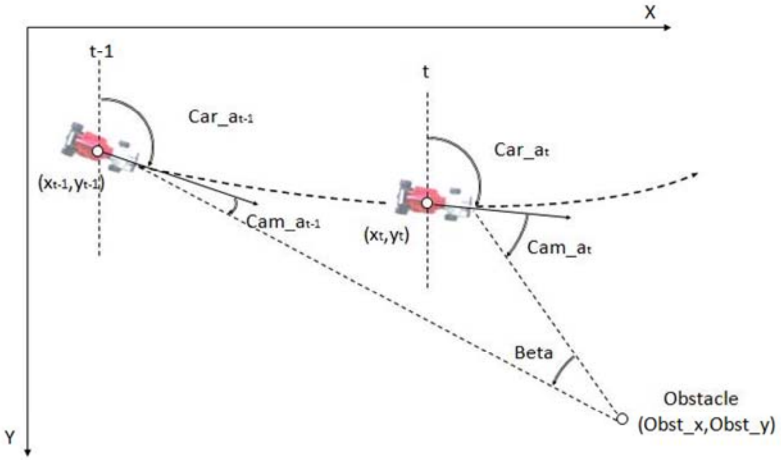


Fig. 5. Defining the global position of the obstacle (Obst_x and Obst_y) by means of fusion of the information about the global position and orientation of the car obtained from the overhead camera (Car_x, Car_y, Car_a, respectively) and the relative position of the obstacle perceived by onboard camera (Cam_a) for two instants of the moving car

3.3 Obstacle Avoidance

Obstacle avoidance is a key feature of any mobile robot [5]. In this very preliminary work we consider the simplest case of a *static* obstacle with *a priori unknown* position, introduced to the scene *during* the time trial. For the considered problem domain of automated control of a scale car, the problem of obstacle avoidance can be viewed as discovering the driving line of circumnavigating an obstacle that result in a safe run around a predefined circuit. The real time detection and mapping of the obstacle, transforms the above mentioned task into a task of dealing with an obstacle with *known* position, introduced to the scene well *before* the initiation of the obstacle avoidance maneuver.

Adopting the repulsive potential field approach of obstacle avoidance, we view the steering the car away from the obstacle as a mechanism of correcting the desired angle of approach (Desired_A_A) of the apex of the following turn. The parameterization of the maneuver is shown in Figure 6. As figure illustrates, the steering correction is initiated when the car enters the obstacle zone, which is assumed to happed *after* the detection and mapping of the obstacle. Consonant with the repulsive potential field approach, the degree of this correction depends on the angular distance between the car and the obstacle (Figure 6, parameter A_O) as follows: the correction A_C of Desired_A_A is set to its maximal (initial) value (Figure 6, parameter A_{CI}) when the bearing of the obstacle is minimal (i.e., A_O=0, when the car travels head on into the obstacle). Then the correction A_C decreases inversely proportionally to zero with the increase of the bearing A_O to its maximal value (i.e., A_O=90 degrees when the car is lined-up with the obstacle). In addition to the steering correction, a throttle control is also applied to maintain the desired velocity V_O while negotiating the obstacle. The obstacle avoidance parameters, predefined by user, are the direction of avoidance

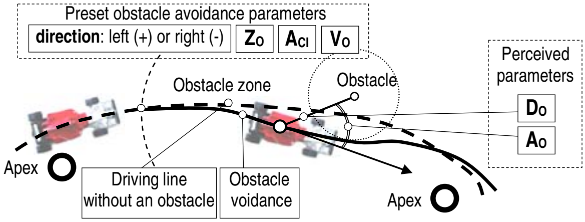

Preset obstacle avoidance parameters
direction: left (+) or right (-)
Zo
ACI
Vo
Obstacle zone
Obstacle
Perceived parameters
Do
Ao
Apex
Driving line without an obstacle
Obstacle voidance
Apex

Fig. 6. Parameterization of obstacle avoidance. Z_O, A_{CI} and V_O are the preset desired values of the radius of the obstacle zone, initial (at the entrance of obstacle zone) correction to the apex approach angle and the speed inside the obstacle zone respectively; D_O and A_O are the *perceived* distance to- and bearing of the obstacle.

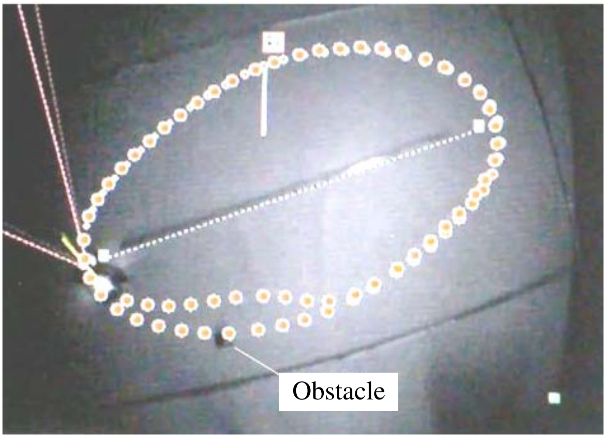


Fig. 7. Driving lines of the car around a predefined circuit before and after the introduction of the obstacle

(left or right), radius of the obstacle zone (Z_O, set to 600mm), initial correction of the apex approach angle (A_{CI} , set to 60 degrees) , and speed inside the obstacle zone (V_O, set to about 1000mm/s).

The driving line before and after the introduction of the obstacle into the scene during the trial, is shown in Figure 7. The experimental results indicate that the average error in the real-time mapping of the obstacle is about 16 cm, which is comparable to the size of the scale car.

4 Conclusion

We demonstrated the feasibility of design of driving agent for automated control of a scale model of a car with obstacle avoidance capabilities. The agent's actions are conveyed to the car via simple remote control unit. The agent perceives the state of the car from live video feed. In order to cope with the inherent video feed latency we implemented an approach of anticipatory modeling in which the agent considers its current actions based on anticipated intrinsic (rather than currently available, outdated) state of the car and surrounding environment. We formalized the notion of driving style with obstacle avoidance capabilities and defined the key parameters, which describe it. The agent perceives static obstacle at unknown position via onboard video camera, maps it in the global scene employing a nature inspired approach based on the motion parallax of the obstacle, and employs a potential-filed approach for its circumnavigation.

References

[1] Abbott, E., Powell, D.: Land-vehicle Navigation Using GPS. The Proceedings of the IEEE 87(1), 145–162 (1999)

[2] Frere, P.: Sports Cars and Competition Driving. Bentley Publishing (1992)

[3] Funge, J.D.: Artificial Intelligence for Computer Games. Peters Corp. (2004)

[4] IBM Corporation: Deep Blue (1997), `http://www.research.ibm.com/deepblue/`

[5] Meeden, L., Kumar, D.: Trends in Evolutionary Robotics. In: Jain, L.C., Fukuda, T. (eds.) Soft Computing for Intelligent Robotic Systems, pp. 215–233. Physica-Verlag, New York (1998)

[6] Robocup (2005), `http://www.robocup.org/02.html`

[7] Rosen, R.: Anticipatory Systems. Pergamon Press, Oxford (1985)

[8] Togelius, J., Lucas, S.M.: Evolving Controllers for Simulated Car Racing. In: Proceedings of IEEE Congress on Evolutionary Computations (CEC 2005), Edinburgh, UK, September 2-5, 2005, pp. 1906–1913 (2005)

[9] Tanev, I., Shimohara, K.: Evolution of Agent, Remotely Operating a Scale Model of a Car through a Latent Video Feedback. Journal of Intelligent Robotic Systems (52), 263–283 (2008)

[10] Wloch, K., Bentley, P.: Optimizing the Performance of a Formula One Car Using a Genetic Algorithm. In: Proceedings of the 8th International Conference on Parallel Problem Solving from Nature, Birmingham, UK, September 18-22, 2004, pp. 702–711 (2004)

Effective Utilization of Neural Networks for Constructing an Intelligent Decision Support System for Dealing Stocks

Norio Baba and Kou Nin

Information Science, Osaka Kyoiku University,
Asahiga-Oka, 4-698-1, Kashiwara City, Osaka Prefecture, 582-8582, Japan

Abstract. In this paper, we propose a new decision support system for dealing stocks which utilizes the predictions (obtained by NNs) concerning the occurrence of the "Golden Cross (GC) and Dead Cross (DC)", those (also obtained by NNs) concerning the rate of change of the future stock price several weeks ahead, and that (also obtained by NNs) concerning the relative position of the stock price versus "GC" and "DC". Computer simulation results concerning the dealings of the TOPIX for the last 15 years confirm the effectiveness of our approach.

Keywords: neural networks, traditional technical analysis, golden cross, dead cross, TOPIX, improved DSS for dealing stocks.

1 Introduction

The widespread popularity of neural networks (NNs) in many different fields is mainly due to their ability to build complex nonlinear relationships between input variables and output variables directly from the training data. NNs can provide models for a large class of real systems which are difficult to handle using traditional approaches [1]-[4].

In this paper, we shall propose a new DSS for dealing stocks which utilizes intelligently the outputs from NNs which give a prediction concerning the occurrence of GC & DC, those which predict increase (decrease) rate of a stock price several weeks ahead, and those which predict the relative position of the future stock price versus crossing point of the two moving averages. The outline of this paper is as follows. In Section 2, we shall briefly touch upon the traditional technical analysis which predicts future tendency of the stock price by taking the relative relationship between "long term moving average" and "short term moving average" into account. Then, we shall mention that a DSS relying upon only the prediction concerning the occurrence of GC&DC becomes sometimes unreliable. In Section 3, we shall propose a new DSS which utilizes not only the predictions concerning the occurrence of GC & DC, but also predictions concerning the rate of change of the future stock price several weeks ahead and those concerning the relative position of the stock price versus predicted crossing point of GC(DC). In Section 4, computer simulation results concerning the dealings of the TOPIX during the rather long range of periods will be given.

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 761–766, 2009.

2 Traditional Technical Analysis in the Financial Market

In order to get an information concerning the current trend of the stock prices and /or indexes such as the TOPIX, the Nikkei-225 and etc., many stock traders have often relied upon the traditional technical analysis which takes the relative relationship between "the Long Term Moving Average (LTMA) " and "the Short Term Moving Average (STMA)" into account. Almost all of them have believed that the Golden Cross (Dead Cross) which STMA cuts LTMA upwards(downwards) gives a strong sign that suggests the upward (downward) moving of the future stock price.

However, recently, we have noticed that GC & DC are not always reliable in making a forecast of future movement of a stock price. Their reliability depends strongly upon the relative changes of the STMA & LTMA near the crosses. Further, quite recently, we have also noticed that their reliability is also strongly influenced by the relative position of the stock price versus the crossing point.

In the following section, we shall propose a new DSS which utilizes predictions (given by NNs) concerning the occurrence of GC (DC), rate of change of the future stock price several weeks ahead, and the relative position of the future stock price versus crossing point of the two moving averages.

Remark 2.1. Fig.1 shows the changes of the Nikkei-225 during each week (candle stick), changes of the short term moving average (solid line; 6 weeks moving average; year 2007), and changes of the long term moving average (dotted line; 13 weeks moving average; year 2007).

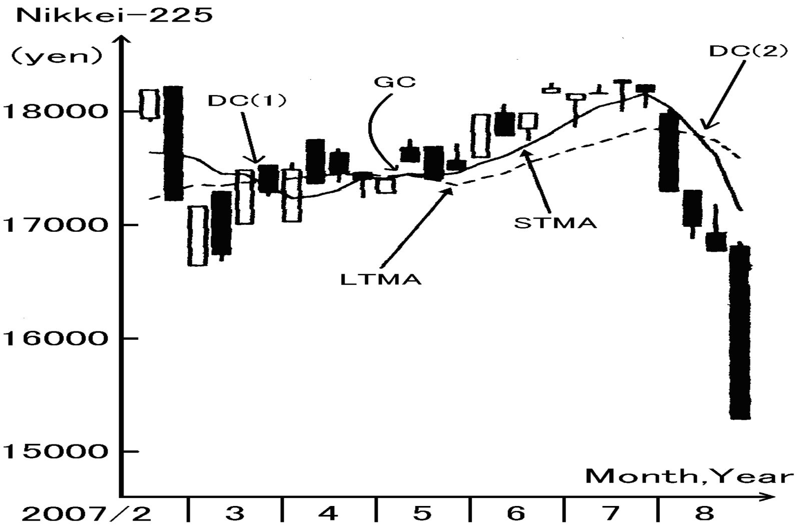


Fig. 1. An Example of the Movements of the LTMA and the STMA near the Dead Crosses of the Nikkei-225

The long term moving average (LTMA) of the Nikkei-225 near the DC(1) does not move downwards. On the other hand, the short term moving average (STMA) of the Nikkei-225 near the DC(1) moves downwards. Further, the price of the Nikkei-225 at the week when the dead cross occurred is quite close to that of DC(1). Nikkei-225 did not go down heavily even after the occurrence of the dead cross.

On the other hand, the situation near the DC(2) is quite different from that near DC(1). The LTMA moves downwards. Further, the STMA moves downwards heavily.Also, changes of the Nikkei-225 occurred far below the crossing point of DC(2).

We can easily observe that the Nikkei-225 has gone down quite heavily after the Dead Cross DC(2).

3 A New Decision Support System (NDSS) for Dealing Stocks Which Utilizes Three Kinds of Predictions Obtained by NNs

In the previous section, we mentioned that effectiveness for utilizing predictions of GC & DC in dealing stocks depends strongly upon the directions of STMA & LTMA near the crossing points and the relative position of stock price versus GC & DC.

In the followings, we shall propose a new DSS which utilizes predictions (obtained by NNs) concerning the occurrence of GC & DC, the increase (decrease) rate of the future stock price several weeks ahead, and the relative position of the predicted stock price versus the predicted crossing point of the STMA & LTMA.

NDSS

Carry out dealing "Buy (Sell)" only when the following three conditions A), B) , and C) are satisfied.

Condition A) : One of the following conditions (A-1) & (A-2) is satisfied.

(A-1) All of the outputs from n NNs are positive (negative) and the average of their absolute values is above 0.5.

(A-2) All of the following three conditions (2-a), (2-b), and (2-c) are satisfied.

(2-a) The greater part of the outputs from n NNs is positive (negative).

(2-b) The rate of the outputs having values over 0.7 (below - 0.7) exceeds 50 % of the number of the outputs having the same sign.

(2-c) The average of the outputs from n NNs is above 0.5 (below – 0.5)

Condition B) : The number of the NNs whose outputs have negative (positive) sign among the m NNs which are prepared for making predictions concerning the rate of change of future stock price several weeks ahead is smaller than (S – 1).

Condition C) : The predicted stock price (by NNs) is above (below) the crossing point of the GC(DC).

Remark 3.1. The n NNs in the Condition A) are prepared for making a prediction of GC(DC) several weeks before it occurs. Due to page, we don't go into details. Interested readers are kindly asked to read the paper [5].

Remark 3.2. Signs "positive" and "negative" in Condition A) and B) correspond to buying and selling, respectively.

Remark 3.3. Condition B) is prepared in order to let the proposed system make no response in the case that many NNs do not show any consistant forecast concerning the changes of the future stock price several weeks ahead.

4 Computer Simulations

We have carried out computer simulations concerning dealings of the TOPIX from 1994 to 2008. Table 1 shows the changes of the initial amount of money (10 billion yen) during each year by utilizing the new DSS (NDSS) proposed in the last section, the PDSS proposed several years ago, After Crossing Method (ACM) which carries out dealing based upon the traditional technical analysis (which carries out dealing only after GC (DC) is found), and the Buy-and-Hold method (BHM) [5]-[7]. The simulation results having been obtained confirm the effectiveness of the NDSS.

Remark 4.1: *In our simulations, we have set n = 7 and m = 21. We have used 7 k×k×1 neural network models (where k denotes the number of input variables having been chosen by the sensitivity analysis [8]) for checking whether GC (DC) will occur in several weeks. We have also used 3×7 = 21 neural network models for making predictions concerning the increase (decrease) rate of changes of each individual stock in the Tokyo Stock Market 3 weeks, 4 weeks, and 5 weeks in the future.*

Remark 4.2: *We have carried out neural network training by using the past data for three years. Table 2 shows the learning periods and the prediction periods.*

Remark 4.3: *In the above simulations, we have used the rule which allows "dealing on credit". Due to space, we don't go into details. Interested readers are kindly asked to attend our presentation.*

Remark 4.4: *In the above simulations, we have taken the charge for dealing into account by subtracting (0.001* (total money used for dealing) + 250,000) yen from the total fund for dealing.*

Remark 4.5: *In the above simulations, we have assumed that we could carry out dealing three times as much as the initial cash "10 thousand million yen" each year.*

Table 1. Total Return in Each Year Which Has Been Obtained by Each Method (TOPIX; Million Yen)

	Buy&Hold	**After Cross**	**PDSS**	**NDSS(s=12)**
A1(1994)	809	-146	681	2,204
A2(1995)	253	452	-3,038	5,862
A3(1996)	-1,029	-266	-298	1,700
A4(1997)	-2,238	-44	9,298	9,230
A5(1998)	-802	-2,071	-1,416	2,091
A6(1999)	5,332	-488	-1,947	7,701
A7(2000)	-2,557	1,037	961	1,358
A8(2001)	-2,055	-66	-683	625
A9(2002)	-2,014	-678	4,646	4,646
A10(2003)	2,432	1,770	11,438	11,438
A11(2004)	764	236	-1,110	-651
A12(2005)	4,365	-479	3,403	6,473
A13(2006)	-43	-1,764	3,765	2,392
A14(2007)	-1,207	-29	2,728	3,182
A15(2008)	-3,924	-683	-3,157	8,960
Total Return	**-1,913**	**-3,221**	**25,270**	**67,211**
	NDSS(s=13)	**NDSS(s=14)**	**NDSS(s=15)**	**NDSS(s=16)**
A1(1994)	2,204	2,204	2,204	2,204
A2(1995)	4,763	4,763	4,763	4,763
A3(1996)	2,680	2,495	2,516	2,516
A4(1997)	7,089	7,089	7,089	7,089
A5(1998)	2,091	2,091	2,091	2,091
A6(1999)	7,701	7,701	7,701	7,701
A7(2000)	1,803	1,803	1,803	1,803
A8(2001)	625	625	625	625
A9(2002)	4,646	4,646	4,646	4,646
A10(2003)	11,438	11,438	11,438	11,438
A11(2004)	-651	-651	-315	-315
A12(2005)	6,473	12,685	12,685	12,685
A13(2006)	4,122	5,046	5,046	4,451
A14(2007)	3,182	4,590	4,590	4,590
A15(2008)	5,407	5,829	1,615	2,245
Total Return	**63,573**	**72,356**	**68,498**	**68,534**

Table 2. Learning Periods and Prediction Periods

	Learning Period	Prediction Period
A1	January 1991 - December 1993	January 1994 - December 1994
A2	January 1992 - December 1994	January 1995 - December 1995
. . .	. . .	. . .
A15	January 2005 - December 2007	January 2008 - December 2008

5 Concluding Remarks

A decision support system for dealing stocks which improves the traditional technical analysis by utilizing NNs has been proposed. Computer simulation results having been done for rather long range of years (15 years) suggest the effectiveness of the proposed DSS. However, these simulations have been done only for the TOPIX.

In order to execute full confirmation concerning the developed NDSS, we need to check whether it can be successfully applied for dealing in the other indexes such as S&P 500, DAX, and etc. For this purpose, we also need to carry out computer simulations concerning various individual stocks. Further, we also need to compare our approach with other approaches such as those utilizing Auto-regressive (AR) linear prediction method [9], and etc. This is also left for our future study.

References

1. Rumelhart, D.E., et al.: Parallel Distributed Processing. MIT Press, Cambridge (1986)
2. Haykin, S.: Neural Networks. Prentice-Hall, Englewood Cliffs (1998)
3. Baba, N., Kozaki, M.: An intelligent forecasting system of stock price using neural network. In: Proceedings of IJCNN 1992, pp. 371–377 (1992)
4. Refenes, A.-P.N., et al.: Neural Networks in Financial Engineering: A Study in Methodology. IEEE Trans. NNs, 1222–1267 (1997)
5. Baba, N., Nomura, T.: An Intelligent Utilization of Neural Networks for Improving the Traditional Technical Analysis in the Stock Markets. In: Khosla, R., Howlett, R.J., Jain, L.C. (eds.) KES 2005. LNCS, vol. 3681, pp. 8–14. Springer, Heidelberg (2005)
6. Baba, N., Nin, K.: Prediction of Golden Cross and Dead Cross by Neural Networks and Its Utilization. In: Apolloni, B., Howlett, R.J., Jain, L. (eds.) KES 2007, Part II. LNCS, vol. 4693, pp. 642–648. Springer, Heidelberg (2007)
7. Baba, N., Nin, K.: Prediction of Golden Cross and Dead Cross by Artificial Neural Networks Could Contribute a Lot for Constructing an Intelligent Decision Support System for Dealing Stocks. In: Proc. of ICCAS 2008, pp. 2547–2550 (2008)
8. Zurada, J.M., et al.: Sensitivity analysis for minimization of joint data dimension for feed forward neural network. In: Proceedings of the IEEE International Symposium on Circuits and Systems, pp. 447–450 (1994)
9. Harvey, A.C.: Time Series Models. Prentice Hall / Harvester Wheatsheaf (1993)

Fine Grained Parallel Processing for Soft Computing

Osamu Fujita and Koji Jinya

Osaka Kyoiku University
Department of Arts and Sciences
4-698-1 Asahigaoka, Kashiwara, Osaka
581-8582 Japan
{fuji@cc,j089607@ex}.osaka-kyoiku.ac.jp

Abstract. This paper considers an approach to fine grained parallel processing for soft computing that mainly deals with large-scale stochastic optimization problems. In the detailed steps of the computation, there are a lot of useless calculations that has no influence upon final results. Removing such a wasted process must be effective to reduce the computational cost. The key is asynchronization of data processing by using redundancy of variables and priority-based processing. A typical system architecture to support this approach is presented and discussed for its application.

Keywords: Parallel processing, Soft computing, Stochastic computation, Asynchronization, Priority-based computing.

1 Introduction

It is very difficult for conventional digital computers to solve large-scale optimization problems such a learning process of very large neural networks and energy minimization for prediction of three-dimensional structure of large organic molecules like proteins. It mainly due to the computational cost that grows exponentially with problem size. Analog computers, instead of digital, might not suffer from this computational drawback, but it must be impractical to produce them artificially. In the present state of the art, there seems no way other than to search quasi-optimum solutions stochastically by using soft computing, even if the computational cost is still high.

In order to reduce the computation time, parallel processing is an inevitable approach. The SIMD model is useful for such an intensive computation as an astrophysics N-body simulation that repeats relatively simple calculations again and again [1]. Its performance can be easily improved by using dedicated hardware. The every step of calculation is perfectly carried out, but this strictness sometimes brings about a waste of time and energy for useless calculations in the above-mentioned large-scale optimization. This situation occurs, for example, when only a small number of variables change state at a time while the other variables are relatively almost stationary. It is worthwhile trying to remove such a useless calculation.

This paper shows an approach to removing useless calculation by using a priority control of fine-grained computation processes. The key is to introduce redundancy (or uncertainty) of variables, i.e., a single variable is represented by a statistic of two or

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 767–772, 2009.

more sub-variables, which enable to relax restrictions on synchronization of data processing and rearrange the order of calculations. An example of the system architecture and its application are shown in the following sections.

2 Target Application Domain

The main application considered hereafter is stochastic computation for successive approximation, iterative simulation, energy minimization and relaxation, especially those having a large number of variables. These computation schemes are often used for solving many kinds of real-world nonlinear optimization problems that can be replaced or approximated by a nonlinear function minimization problems.

The noteworthy characteristics of these problems are as follows:

1) The system involves a large number of variables (or parameters).
2) A connection matrix that represents interactions or relationships between all variables, in which each element may be a simple numerical value or a complicated operator, should be treated as a dense matrix in general, though it can be approximated to be sparse in practice. The connection strength may dynamically changes during the computation.
3) In the process of searching for a solution, which can be represented by a trajectory of a discrete evolution equation, the system is almost convergent, i.e., the final result of computation is insensitive to some error or noise introduced in the calculations.

The third condition is restrictive for the application, but there are many practical problems that meet this condition in cases, for example, where the objective function and/or its variables are statistical or probabilistic, and the model of the system intrinsically includes non-trivial noise.

3 Parallel Processing for Soft Computing

Parallel processing system architecture can vary depending on the type of the computational model. Choosing an appropriate architecture for the above-mentioned problems is important issue for improving cost performance.

3.1 Conventional Fine Grained Parallel Processing

For the case that the computational model of problem solving includes large matrix operation, it is necessary for parallel processing to partition the matrix operation, however there is no appropriate way of partitioning of dense matrix operation into coarse grained processes. It should be better to resolve into fine grained processes such as primitive calculations of each variable or matrix element. Hence, rather than using Grid computing, it might be more appropriate to use a systolic array and/or SIMD architecture. These architectures are intended to execute deterministic and strict calculations parallelly, and so they generally require perfect synchronization of data processing. There is no room to introduce such an interruptive control as removing useless calculations and rearranging the order of computation processes.

3.2 New Approach to Fine Grained Parallel Processing

Putting the above-mentioned characteristics of the soft computing to use, it enables to implement the interruptive control in fine grained parallel processing. There are two points: asynchronization of data processing based on the redundancy of variables and priority-based fine grained task control.

The redundancy of variables is realized by substituting a single variable to such a statistic as mean, median and mode of two or more sub-variables. A processer executes a program computing with the representative values of sub-variables and updates only one sub-variable at a time. Two or more processors may execute each program asynchronously, no matter whether the programs are the same or not.

Updating sub-variables can be executed independently of each other so that the order of execution may be changed according to priority. In case that some variables have been changing greatly and the others are relatively stationary, it should be better to execute updating with respect to the former dynamic variables before updating the other stationary variables that are already converged to a final state or temporarily ineffective and useless for convergence. The priority of the task can be defined by the degree of dispersion of relevant sub-variables in this case, because an unconverged variable probably fluctuates so that its sub-variables are scattered accordingly. The priority can change dynamically depending on the situation of variables in convergence process.

Although the redundancy of variables and the priority control have negative effect to increase the memory space and computation time, it is expected the total computational cost can be reduced by removing useless and wasteful calculations.

4 System Architecture

An example of the system architecture is shown in Fig. 1. The system is composed of a shared data memory, one or more processing elements, and a task controller. In the shared data memory, every variable is stored together with a set of its sub-variables and their statistics such as mean for the representative value and variance for the priority factor. The task controller has a priority list of variables and outputs the index of the highest priority variable in response to the request from the processing elements. Each processing element executes its own programs stored in each local memory, only with respect to one or more variables highly relevant to the highest priority variable indicated by the task controller.

The priority list has data set of the index of variables and its priority value, where it is better not to have duplicate index and unnecessary to have complete list of all variables. This list is updated when the highest variable is read out and erased from the list and new priority data are received from the processing elements. The function can be implemented in either hardware or software, which depends on the computation time per task, the numbers of variables and sub-variables, and the numbers of processing elements and task controllers. Two or more task controllers may be used and connected to the processing elements with a variety of network topologies. For the fastest updating of the list, in hardware implementation, the best performance can be obtained by using a register array where each register has own data comparator and

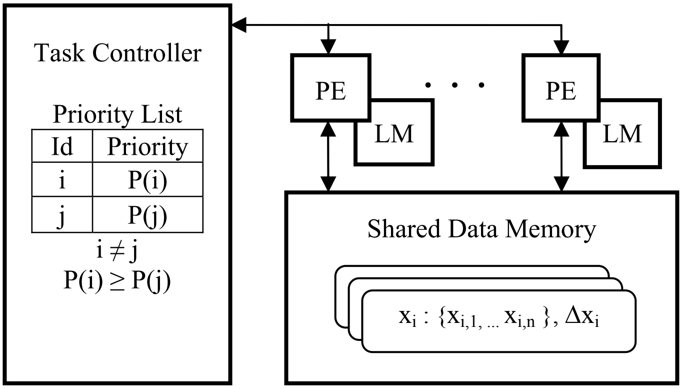


Fig. 1. An example of the system architecture: The system composed of a Shared Data Memory, Processing Elements (PE), each having its own local program memory (LM), and Task Controller

selector. The register array stores the priority data according to the order of priority values, and can execute updating the priority list every clock cycle.

Each processing element reads out the index of the highest priority variable from the task controller, executes tasks highly relevant to the variable indicated by the index, and updates one or more sub-variables, representative values of updated sub-variables and their priority values in the shared data memory, independently from each other except for some restrictions such as an anti-collision control of write operations. In updating of a set of sub-variables, it is better for the replacement to select a sub-variable whose value is the farthest from the new value to be updated. If the priority values are changed, the processing element sends a message to the task controller to update the priority list.

5 Application Example

As a simple example, let us consider a fully connected neural network that can be used for solving a variety of optimization problems [2]. The solution is obtained by the neural network simulation based on the discrete-time approximation of the equation of motion that can be represented by

$$u_i(t+1) = u_i(t) - a\frac{\partial}{\partial u_i}\left(\frac{1}{2}\sum_{j=1}^{N}\sum_{k=1}^{N} w_{jk} f(u_j) f(u_k)\right) \tag{1}$$

where u_i is the i-th variable, N is the total number of variables, t is a discrete time step, a is a parameter to control simulation performance, w_{jk} is the element in the j-th row and k-th column of the connection matrix $\mathbf{W}$, and $f()$ is a nonlinear function of the variable.

The connection matrix is usually symmetric so that the system has a potential energy function in the quadratic form. The iterative calculation of Eq. 1 leads to a solution as a local minimum state of the energy function. There is generally no restriction with respect to the magnitude of off-diagonal elements, but there might be some natural patterns of the distribution of connection values, such as power law or log-normal distributions, for a very large matrix. If the matrix can be approximated by a sparse matrix by emphasizing the strong connections, the value change of a variable affects only a small number of connected variables at a time. In such a condition, regarding the calculation of Eq. 1 for one variable as a single task, it is worthwhile to apply priority based task control so as to calculate only the small number of variables that expected to be changed their values.

The function f is generally defined as a sigmoid function having a large saturated region where the first derivative is negligibly small. When u_j is in the saturated region, the change of $f(u_j)$ can be very small even if the change of u_j is very large. Thus, there must be a lot of variables that are approximately stationary no matter they have not converged to a final stationary state. The computation tasks necessary to be executed can be selected taking into account the sensitivity to changes of another variable, which depends on and varies with the situation of variables.

In fact, for example, in the software simulation of the Hopfield neural network to solve traveling salesman problems [2], detailed analysis of the trajectory of variables has revealed the existence of lots of useless and wasteful calculations that the change of the variable is too small to contribute to energy minimization, i.e., the final result is almost the same, no matter whether such a small change is taken into account or ignored. Therefore, using the system shown in Fig. 1, regarding u_i and/or $f(u_i)$ as variables and assigning the maximum of difference between their sub-variables as their priority value, the priority based task control will be effective in removing the useless calculations and reducing the computational cost. As the number of variables increases, the effect would be expected to increase.

6 Future Problems

It is important for performance evaluation to implement the system in massive parallel hardware. This is mainly because, if the number of variables is not large enough, it is difficult to evaluate the effectiveness of the proposed approach. Software simulation is generally suffering from the limitation of the problem size due to concerns about computational complexity and is difficult to use for practical applications, which is just the reason why we investigate the new approach. Another reason is to reduce not only the computation time but also wasted power consumption. The fine-grained tasks may be relatively simple numerical calculations so that the processor element can be made simply. It must be better to make processors as small as possible and massively integrate them on a single chip using FPGA or PLD.

The stochastic computation uses some parameters, functions and procedures for performance control that affect convergence of the solution. In the proposed system, for example, a function for determining the priority value, its reevaluation procedure, and the timing control of when to update the priority list will play significant roles in the task control. However, unfortunately, there is neither adequate theory nor practical knowhow at present.

In application for real-world problems such as 3D structure prediction of proteins, the computational model might have to be designed somewhat complicated so as to simulate efficiently according to the hierarchical structures of natural molecules.

7 Concluding Remarks

In order to overcome difficulties in conventional software simulation for soft computing, a new approach of parallel processing has been considered. The redundancy of variables using sub-variables can be associated with the particle filter that is used for the Markov Chain Monte Carlo method [3]. Reducing the computational cost, i.e., removing useless and wasteful calculations by using priority based fine grained task control, might be called "processing compression" in contrast with "data compression" that removes meaningless data. We hope to achieve a very high compression rate as a trade-off with an allowable loss of accuracy.

References

1. Sugimoto, D., Chikada, Y., Makino, J., Ito, T., Ebisuzaki, T., Umemura, M.: A special-purpose computer for gravitational many-body problems. Nature 345, 33–35 (1990)
2. Hopfield, J.J., Tank, D.W.: Neural Computation of Decisions in Optimization Problems. Biol. Cybern. 52, 141–152 (1985)
3. Doucet, A., Johansen, A.M.: A tutorial on particle filtering and smoothing: fifteen years later. Technical report, Department of Statistics, University of British Columbia (December 2008), http://www.cs.ubc.ca/%7Earnaud/doucet_johansen_tutorialPF.pdf

New System Structuring Method That Adapts to Technological Progress of Semiconductors

Kunihiro Yamada, Kouji Yoshida, Masanori Kojima, Tetuya Matumura, and Tadanori Mizuno

Tokai University 2-2-12, Takanawa, Minatoku, Tokyo, 108-0074, Japan
yamadaku@tokai.ac.jp
Shonan Institute of Technology,
Osaka Institute of Technology, Renesas Technology Corp. Sizuoka University

Abstract. In the last half century, systems such as those incorporated in electronic equipment and their constituent semiconductors coexisted favorably and supplemented each other while individually performing their own duties. In the 1990s, while low-price semiconductors with yet higher performance than those required by the systems became available, the market however started to see critical quality problems and the manpower and time required for system development increased. These problems are considered to be a result of the improper use of computer-aided design (CAD) that kept pouring an abundant supply of semiconductors into hardware logic circuits and program codes while failing to design what the systems should be like.

Presenting two systems which the authors are currently developing, this paper proposes a new method of designing effective systems while minimizing the costs necessary for system development, production and operation.

Keywords: System, semiconductor, system configuration, design, hardware, and software.

1 Introduction

In the early 1990s, low price yet high performance semiconductors became available for engineers who were newly familiar with designing hardware and software associated mainly with microprocessors. As a result, backed up by powerful CAD, they further expanded system development. If the system configurations were poor or even wrong, the CAD and CAD-generated software were able to make the systems operate. There were no true system engineers but CAD operators pretending to be system designers. As a result, being indifferent to what the system should be like, hardware and software designers stopped cooperating with others while indulging in production within their own confinements. This made the systems even more complicated than necessary while causing quality problems and increasing development manpower and time.[1]

Here, the authors propose a new system configuration method that minimizes the lifetime costs necessary for system development, production and operation. Even when they have the same functions, two systems with different production or

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 773–781, 2009.

operation conditions can end up with different system configurations. To represent this phenomenon, this paper examines two systems from the authors' research. One of them requires further series developments. The other particularly requires adjustment during its operation.

Anticipating that semiconductors will develop and improve even further, the following part of this paper reiterates that systems and their constituent parts once coexisted and performed their own share of duties in an attempt to verify the proposed new system configuration method.

2 Verifying That Systems and Semiconductors Once Coexisted

Systems and semiconductors coexisted and supported each other over the period from the second half of the 1960s to the 1990s when semiconductors started to make a significant contribution to systems. Since then however, the market has witnessed the failure of the system design up to the present. On the other hand, important semiconductor-related inventions have continued up to the present day since the discovery of the semiconductor's wave detecting function in 1874 and the invention of point-contact transistors and junction transistors in 1947 and 1949 respectively.[2,3]

The wafer process, which is one of the most important semiconductor technologies, is expected to grow for another ten years so that one chip is expected to integrate as many as 100 times the present number of transistors or 10^{10} transistors as shown in Figure 1.[4] This is a viewpoint that makes our new system configuration method critically important.

Figure 2 shows the system developments in relation to videotape recorders (or VTRs) from 1976 to 1996 when the systems and their parts (semiconductors) still coexisted and supported the other while performing their own duties.[5] The repeated

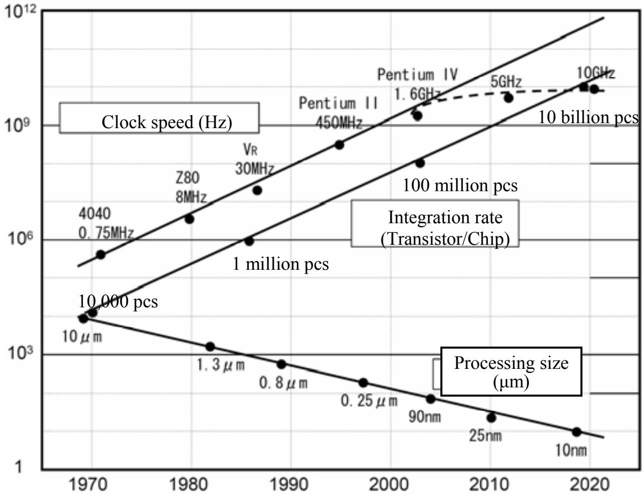


Fig. 1. Contribution of processing ultra-fine wafers to semiconductor integration and clock speeds

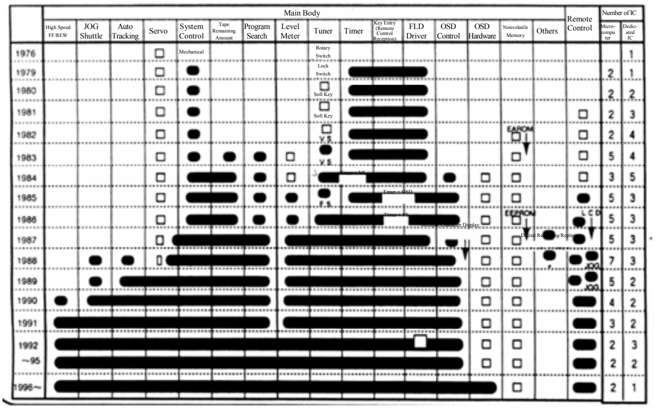


Fig. 2. Improved VTR functions and integration of LSIs

addition and integration of LSIs expanded the VTR functions. A revolutionary change also took place in the mode of processing from analog to digital. All these led to reducing the number of LSIs to as few as one or two. A 100 times increase in the number of semiconductors in the past 20 years has translated into expanded functions, improved performance characteristics and cost reduction of VTRs as well as the revolutionary change in the mode of processing. With a great deal of effort, knowledge and wisdom, designers then carried out all these innovative tasks in a manner and with pride that true designers should live up to.

Figure 3 shows the integration of LSIs. Twelve LSIs for the memory, microprocessors and the logic circuits are integrated into one LSI. It should be noted that integration into one single LSI brings about several benefits. One is a reduced number of

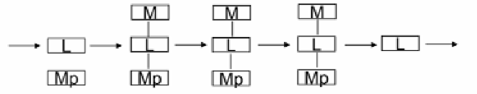
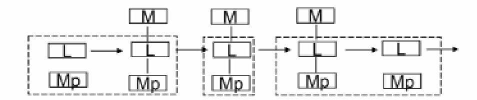
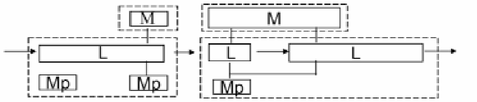
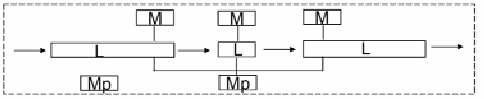
	Configuration of System LSI and Element (M = Memory, L = Logic, Mp = Micro-Processor)	Number of LSI	Number of process block
(a)	 M M M L L L L L Mp Mp Mp Mp	M 3 L 5 MP 4 Total 12	3 5 4 12
(b)	 M M M L L L L L Mp Mp Mp Mp Mp	M 3 L } 3 MP Total 6	3 5 4 12
(c)		M 2 L } 2 MP Total 4	2 3 3 8
(d)		M L } 1 MP Total 1	3 3 2 8

Fig. 3. Reduction in number of LSIs due to development of system LSIs

input/output signals communicated with peripherals through the LSI terminals. This-makes the LSI internal configuration more natural. It relates to the logic circuits and memory for data compression and decompression in image processing. The number of microprocessors is reduced from three to one.

This section of the paper presents an example of the VTR's motor-driven servo circuit to discuss the revolutionary change in system processing.[6] Figure 4 shows that the analog servo circuit was digitalized around 1980 and was integrated into a microprocessor in the middle of 1980. It changed further to a software servo circuit from around 1990 onwards. The digital servo benefits automatic pre-shipment adjustment and improves the performance characteristics. The software servo can automatically adjust changes due to variations in the motor characteristics and a secular change through learning. It is even more significant that the number of necessary logic circuits is reduced and that logic circuit design replaces program design to reduce the manpower and time for system design.

Technological improvements in the servo circuits contribute to these innovative developments. The increase in the number of semiconductors on a single chip and the development of high-speed microprocessors, both of which have resulted from the micro-technology of processing ultra-fine wafers, also have played an important role in servo circuits.

The technological innovations from analog processing to digital processing and further to software (program) processing are typical of the system growth and improvements in the VTR's motor drive system described above. It is considered that this suggests the future of system processing.

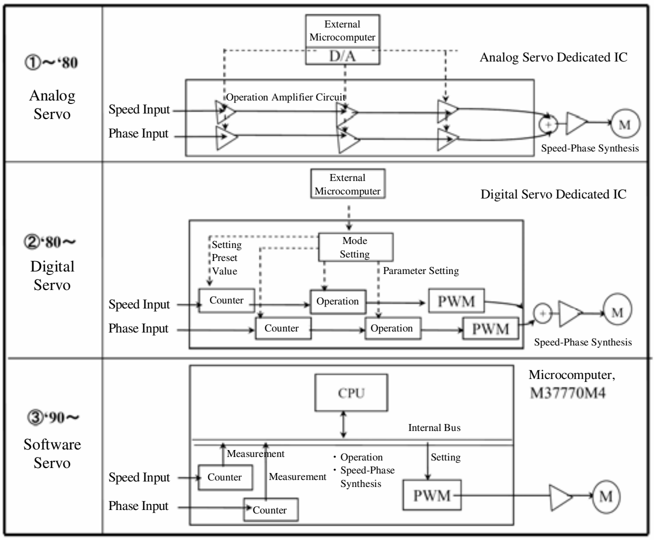


Fig. 4. Improvement in motor servo processing systems

3 New System Configuration Method

This paper defines the new system configuration method as a method of creating systems while minimizing the sum of the system lifetime costs for development, production and operation, which also determine the system configuration. Here, the operation cost refers to the cost for installation and adjustment to make the system operable. The concept of the three lifetime costs indicates the realization of a new system whose configuration the market has never seen before.

A system configuration that minimizes the sum of all costs is mostly and quite naturally derived by partially modifying software or hardware that is already on the market or developed previously. Applying the new system configuration method every time when developing a new system helps avoid the potential pitfalls resulting in unfavorable systems in the absence of true design and provides the market with effective systems. The following section of the paper discusses two cases in which, in the authors' judgment, such partial modification or remaking of the hardware or the software of an existing system does not apply.

One of them concerns a digital TV or DVD capable of MPEG compression and decompression of image and audio data. Since such items involve a series of developments, a microprocessor with a second ALU deals with it. The other is a network system such as one for home use that characteristically calls for a lot of adjustment and correction when operating the system. Two or more means of communication handle the second case.

4 Microprocessor with Second ALU

A system capable of MEPG compression and decompression of digital TV images and audio signals requires 60 hardware circuits to cover all possible combinations of broadcast regions, number of image pixels and other particulars. More specifically, the 60 different modes of broadcasting are derived by combining three regional systems (NTSC, PAL and SECOM), five different numbers of pixels (HDTV, SD, conventional analog, and VGA and SXGA for personal computers), two types of scanning (interlace and progressive), and two types of screen size ratio.

A dedicated processor called a second ALU which is independent from the main processor bus program controls these 60 different types of MPEG compression and decompression. Compared with the development of a system involving 60 different MPEG hardware circuits, a system with a single hardware circuits plus 60 programs can drastically cut down on the development manpower and time to an estimated 216 man-months (or 23% of 936 man-months that would otherwise be required).

A dedicated processor has the potential to become smaller than the comparable random logic. The authors examined a method in which a decentralized system having a second ALU and bus switches independent of the processor carries out MPEG2 IDCT processing. The system consists of a small circuits made up of a 1-Kbyte gate logic and a 2-Kbyte RAM using about 20,000 transistors and of a small-scale software program of only 13 steps.[7]

Figure 5 shows the structure of the second ALU. The address space for the second ALU can be very small.

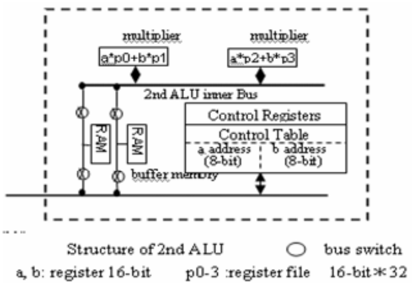


Fig. 5. Schematic configuration of second ALU

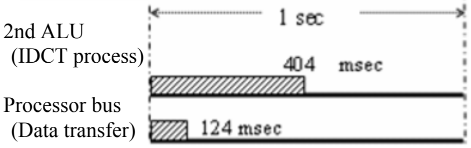


Fig. 6. Buffer memory processing time for each second

The program responsible for IDCT processing in MPEG2 is as small as 13 steps. The decentralized system has bus switches that make the second ALU independent within a 32-bit microprocessor. It also has a 2-Kbyte data memory, a 4-sum-of-product ALU and a control register.

The authors evaluated the time necessary for the IDCT processing and data transfer at a clock bus rate of 250 MHz. The IDCT processing took 13.3 microseconds to process one buffer memory (or eight blocks). The second ALU is driven for 404 microseconds for one second.

Data transfer between the microprocessor and the second ALU took 4.1 microseconds where the buffer memory was configured to 32 bits × 256 words while using the DMA for readout after the IDCT processing and data setting. As shown in Figure 6, the process was executed 30,375 times in one second while occupying the processor bus for 124 microseconds.[8]

5 Mutual Supplement Network between Wired and Wireless Systems

If one communications system is not enough for the network to develop its intended communications performance, multiple communications systems should be used rather than attempting to improve that particular single communications system. More specifically as shown in Figure 7, a network of dual communications systems is organized by adding another communications system with different characteristics. The dual communications system sends out data from node A to node B through both wireless (Ld1) and wired (Wd1) lines simultaneously. The communication is successful if one of them works. In the example shown in Figure 8, the wireless and wired communications systems are 82% and 76% successful respectively. When these two support with the other, the success rate increases to 96%.[9] The authors call this particular system a ‘wireless and wired mutually supplementary network’.

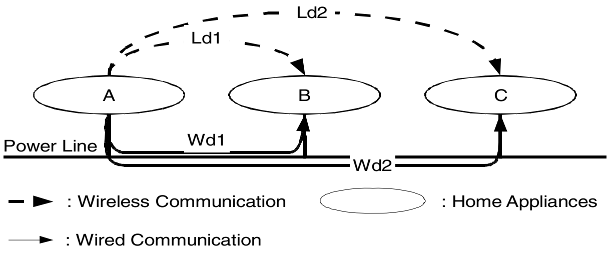
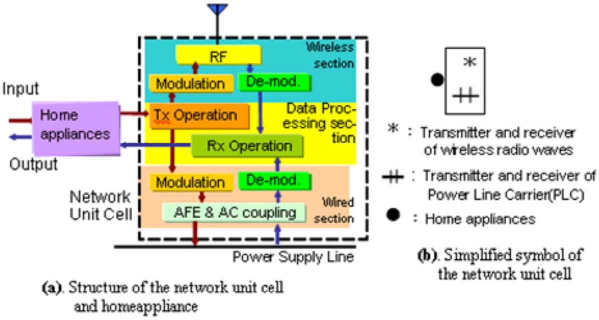


Fig. 7. Wireless and wired mutually supplementary network

		Wireless		
		Truth. 82	**Truth? 14**	**False. 4**
Wired	**Truth. 70**	**57.4**	**9.8**	**2.8**
	Truth? 10	**8.2**	**1.4**	**0.4**
	False. 20	**16.4**	**2.8**	**0.8**

Fig. 8. Comparative aspect of communications quality



(a). Structure of the network unit cell and homeappliance

(b). Simplified symbol of the network unit cell

Network Unit Cell

Fig. 9. Node configuration to realize mutual supplementation (network unit cell)

The wired communications system uses power line communication (PLC) while the wireless uses Zig-bee as specified in IEEE802.15.4. Figure 9 shows the node configuration of the network unit cell of the wireless and wired mutually supplemented network. It consists of three sections that are responsible for wireless communication, wired communication and data processing respectively.

6 Discussions

In this section, this paper compares the data processing capabilities of microprocessors including a digital signal processor (DSP) in a DVD's backend SOC (system on a chip) and the logic controller in an attempt to determine the future aspects of system configuration methods. The data processing efficiencies of these two are calculated by dividing the data memory capacities under the control of the microprocessor and the logic

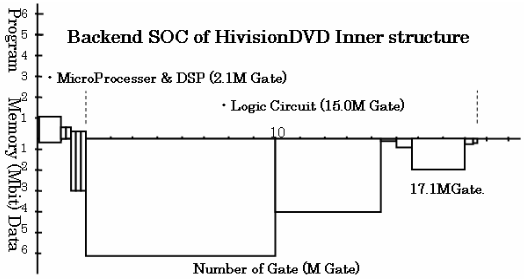


Fig. 10. Numbers of gates and memory capacities at a DVD's backend SOC

controller by the number of gates of the microprocessor and the number of the gates of the logic controller respectively. A higher calculated result means higher efficiency.

Figure 10 shows the numbers of gates and the memory capacities of the microprocessor for the DVD's backend SOC and the logic controller. The number of gates are shown on the horizontal (x) axis. The program memory capacities are shown on the vertical (y) axis above the horizontal (x) axis and the data memory capacities are on the vertical axis below the horizontal (x) axis.

The numbers of gates (G) for the microprocessor and the logic controller are about 15.0 mG and 2.0 mG respectively being 17.0 mG in total. The microprocessor dictates an 8.3 mB data memory while the logic controller dictates a 14.0 mB data memory. The resulting data processing capability of the microprocessor is 4.2 GB while it is 0.9 GB for the logic controller, indicating that the microprocessor uses the logic gates 4.7 times more effectively than the logic controller. In an extreme case, if the microprocessor is assumed to replace the entire logic controller, the total number of gates at the SOC could be reduced to 5.2 mG or 30.1% of the current 17.0 mG. The authors will later verify this assumption and specific replacement of the logic controller by the microprocessor.

7 Conclusions

Presently, both hardware and software are unnecessarily complex, adversely affecting quality and the development man-months count. The authors have proposed a new system configuration method to rectify this unfavorable trend. Defining it as a method of creating systems while minimizing the sum of the system lifetime costs for development, production and operation, which also determine the system configuration, the paper discussed two system examples. The system configurations of these two differ significantly from each other. One of them has a processor called a second ALU and replaces the logic control circuit with program processing so as to reduce the development manpower and time and the number of gates. The other is a wireless and wired mutually supplementary network, a combination of a wireless and a wired communications system, which substantially improves the communications performance. While the first example reduces the number of gates, the second one on the other hand increases the number of gates. The total cost, in any event, should be the governing factor in determining the system configuration.

Furthermore, as discussed earlier in relation to a DVD's backend SOC, the authors will later examine the characteristics and possibility of the microprocessor and the logic controller to verify the effectiveness of replacing hardware with software.

References

1. Nikkei Electronics Embedded Software, pp. 22–24 (2006) ISBN4-8222-0249-2
2. Braun, F.: Uber die Stromleitung durch Schwefelmetalle. Ann. phys. Chem. 153, 556 (1874)
3. Shockley, W.: The Theory of p-n Junction in Semiconductor and p-n Junction Transistor. Bell Syst. Tech. J. 28, 435 (1949)
4. ITRS (International Technology Roadmap for Semi-conductors), 2003 edn., SIA
5. Yamada, K., et al.: Applied Technology of One-chip Microcomputer and Software. Mitsubishi Electric Technical Review 66(2), 229–235 (1992)
6. Hayashi, K., Yamada, K., et al.: 16-bit Microcomputer for VTR Software Servo. ditto, developed based, pp. 205–213
7. Yamada, K., et al.: A New RISC Processor Architecture for MPEG-2 Decoding. IEEE Trans. On Cons. Elec. 48(1), 143–150 (2002)
8. Chen, W.A., et al.: A Fast Computational Algorithm for the Discrete Cosine Transforms. IEEE Trans. on Communications COM-25(9), 1004–1011 (1977)
9. Yamada, K., et al.: Dual Communication system using wired and wireless correspondence in a small space. In: Negoita, M.G., Howlett, R.J., Jain, L.C. (eds.) KES 2004. LNCS (LNAI), vol. 3214, pp. 898–904. Springer, Heidelberg (2004)

A Network Approach for HIV-1 Drug Resistance Prevention

Kouji Harada and Yoshiteru Ishida

Toyohasi University of Technology,
1-1, Tenpaku, Toyohashi-shi, Aichi 441-8585, Japan
{harada,ishida}@tutkie.tut.ac.jp

Abstract. In AIDS treatments, it is an imperative problem to reduce the risk of the drug resistance. The previous study discussed which HIV-1 gene products are an ideal drug target not to develop drug resistance by applying some ideas of the graph theory, and suggested that the drug resistance would not develop if the drug target molecule functions as "hub" in a chemical network where HIV-1 gene products interact directly or indirectly with intracellular agents in a HIV-1 host cell. The present study fortifies this suggestion in mathematical framework. The study develops the expression for a probability of drug resistance developing over the two different types: non-hub and hub of drug targets, and demonstrates that the hub drug target is more favorable for the drug resistance prevention than the non-hub one.

1 Introduction

HIV-1 is a causal agent of AIDS and belongs in retrovirus family of RNA viruses that synthesize a DNA copy of their RNA genome after infection of the host cells. HIV-1 host cells are CD4^{+} lymphocytes or macrophages [1].

Now we have only drug treatment against AIDS. In Japan, 22 anti-HIV drugs were approved by 2008 [2]. Almost the existing anti-HIV drugs are categorized into 3 classes: HIV reverse transcriptase inhibitors, HIV integrase inhibitors and HIV protease inhibitors.

HIV reverse transcriptase inhibitors (RTIs) are categorized into two sub classes: nucleotide analog reverse transcriptase inhibitors (NRTIs) and non-nucleotide reverse transcriptase inhibitors (NNRTIs). NRTIs act as a nucleotide analog and to arrest DNA chain elongation by HIV-1 reverse transcriptase. As representative NRTIs, abacavir (ABC), zidovudine (AZT), lamivudine (3TC), stavudine (d4T), etc are popular. On the other hand, NNRTIs act as non-competitive antagonists of enzyme activity by binding to the catalytic site of the HIV-1 reverse transcriptase. Representative NNRTIs are nevirapine (NVP), efavirenz (EFV) and delavirdine (DLV).

HIV integrase inhibitors (INIs) block a HIV integrase enzyme to integrate a HIV DNA into a host cell's DNA. INI is a relatively recently developed anti-HIV drug. In Japan, only Raltegravir (RAL) is approved.

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 782–788, 2009.

HIV protease inhibitors (PIs) block HIV protease's catalytic activity that cleaves HIV Gag-Pol polyprotein into HIV functional enzymes such as a reverse transcriptase, a protease, an integrase and HIV structural proteins like matrix proteins and capside proteins. PIs are categorized into two subclasses: hydroxyethylamine isosteres and symmetrical inhibitors [3]. The hydroxyethylamine isosteres is an analogous of (Tyrthyrosine/Phenilalanine)-Proline segments, which is a cleft site of the Gag-Pol polyprotein. As representative hydroxyethylamine isosteres PI, saquinavir (SQV) and nelfinavir (NEV) are well-known. Also, the symmetrical inhibitor is designed as its P1-P1' segments to be symmetrically positioned to a S1-S1' segment in HIV PR catalytic site. As the symmetrical inhibitor PI, ritonavir (RTV) and lopinavir (LTV) are provided.

These antiretroviral agents have the drug resistance problem. Drug resistance mutations arise as a base substitution of when HIV reverse transcriptase transcribes a HIV RNA genome into a HIV DNA copy. On the listed drugs, resistance mutations and cross-resistance were observed. That is why it is important to figure out some undiscovered attributes of anti-HIV drug targets to produce less resistance mutations than the existing ones. Our previous study proposed "hub-ness" as an attribute of drug targets not to lead to developing drug resistance [4]. The concept of hub-ness is introduced in the graph theory to represent a characteristic node having links to the other many nodes in some network. Here hub drug target means the drug target has direct binding interaction with many intracellular agents. The present study estimates probability of a non-hub/hub drug target developing the resistance and demonstrates hub drug target is harder to develop drug resistance than non-hub one.

2 Network Approach for the Drug Discovery

Currently the drug discovery method considering knowledge from biological network research field is gaining recognition. The present dominant drug discovery method is to design a ligand to act on individual drug target. However, growing number of experiments and theoretical works which make pharmacologists doubt the dominant method in terms of effectiveness are appearing. Genomics studies of many model organisms have revealed single gene knockouts little affect phenotype [6,7,8]. Those studies suggest that many single gene functions are redundant, thus a drug acts on a single gene product may be less effective. The redundancy of the gene functions is due to a scale free structure of the network because in general the scale free network is robust to random node declinations corresponding to gene knockouts [9,10]. In order to modulate a robust network, it needs simultaneous modulations of multiple nodes. In facts, simultaneous dual gene knockouts have shown synthetic lethality, synthetic sickness; the isolated knockout of the two individual genes have shown no effect [11]. In antiboitics treatments, many effective antibiotics such as β-lactams, fluoroquinolone and D-Cycloserine etc. target simultaneously multiple proteins rather than individual proteins [12,13]. A series of these demonstrations and discussions provide pharmacologists enough evidences to question the single target drug paradigm.

Hopkins points the importance of designing a compound binding to multiple drug targets and proposes a new concept of network pharmacology [14].

The two challenges the network pharmacology facing are developing methods (1) to identify a combination of nodes to produce a biological network perturbations leading to a therapeutic effect; (2) to find a drug agent to perturb those nodes. For the challenges, now three methods: systemic screening, knowledge-base approach and network analysis, are though to be promising [14].

The systemic screening, such as combination screening of mixtures of drugs is efficient to systematically discover new drug-drug combinations and synthetically lethal gene pairs. However unfortunately combinations discovered in the laboratory do not necessarily apply to the clinic [15]. It is also a troublesome problem that the global search space for the combinations is too vast to reduce it.

The Knowledge-base approach (associate method) enables pharmacologists searching for a new drug therapy through the mixtures of the existing therapies. A successful example is the highly active antiretroviral therapy (HAART) for AIDS treatments. However, this approach has a drawback of not to be able to explain counterintuitive, paradoxical and unexpected network system responses against some medical perturbations.

Network analyses provide pharmacologists a meaning to compensate the drawback of the knowledge-base approaches. Recent network analysis studies suggest network structures relate with emergence of protein functions. For instance, (1) a hub with high betweenness has pleiotropic functions across the network [16]; (2) a bottleneck with high betweenness correlates with the gene expression dynamics [17]; (3) non-hub bottlenecks with transient interactions [18] and bridging proteins [19] are less lethal than average and tend to be independently regulated.

As shown in this section, theoretical and experimental challenges for the establishment of the network pharmacology are initial yet, and still such approach is a minority in the pharmaceutical industry. However it is also true that the concepts of the network pharmacology is coming to the front. The present study takes the network approach for HIV-1 drug target discovery.

3 Hub-Ness of HIV-1 Drug Targets

This section takes up protease (PR), integrase (IN) and reverse transcriptase (RT) as HIV-1 drug targets, and shows that these drug targets are categorized into two types from viewpoints of hub-ness.

Ptak et al estimated number of intracellular agents binding physically to each of 15 HIV-1 proteins [5]. The first column of the table 1 is data on PR, IN and RT among them. The second column shows the total amino-acid residues of each enzyme. Table 1 shows RT is a relatively large protein to IN and PR, but nonetheless RT interacts only two cellular agents, so RT is regarded as a non-hub drug target. On the other hand, IN and PR interacts more than 60 intracellular agents thus both of the enzymes are regarded as a hub drug target.

Table 1. Hub-ness of each drug target

Drug target	*The number of cellular agents*	*The number of amino-acid residues*
RT	2	984
IN	64	288
PR	63	198

To sum up, drug targets are categorized into at least 2 types.

- Non-hub drug target
- Hub drug target

The next section estimates a probability of the two types of drug target developing drug resistance, and demonstrates which type of drug targets is favorable for the drug resistance inhibition.

4 Hub-Ness Favorable for the Drug Resistance Inhibition

This section estimates a probability of each of the two types: non-hub type and hub one, of drug target developing resistance to an antiretroviral agent. For simplicity, this study makes the following assumptions.

- Drug target interacts with a drug and k ($k \geq 0$) intracellular agents (especially, $k = 0$: non-hub)
- Binding between a drug and its target is dominated by one amino-acid residue within the drug-binding site on the target.
- The transcriptional error rate of RT is constant with an amino-acid residue position of a drug target.
- Any amino-acid substitution happening out of drug- or intracellular agent-binding sites are irrelevant with binding between a drug and its target, or intracellular agents and their targets.

In the following sections, let $P(X_i \rightarrow X'_i)$ to represent a probability of amino-acid X_i mutating into amino-acid X'_i at an amino-acid residue position i.

4.1 Non-hub Drug Target

This subsection treats with a non-hub drug target which does not have any interaction with intracellular agents. It supposes that a drug binds with an amino acid residue X_L at a position L on one-dimensional amino-acid sequence of a drug target protein, and the drug target develops resistance to the drug when an amino acid X_L mutates into X'_L (Fig. 1). Let the drug resistance probability of non-hub drug target to represent P_{nh}, then the next relational expression succeeds.

$$P_{nh} = \sum_{X'_L(\neq X_L)} P(X_L \rightarrow X'_L), \tag{1}$$

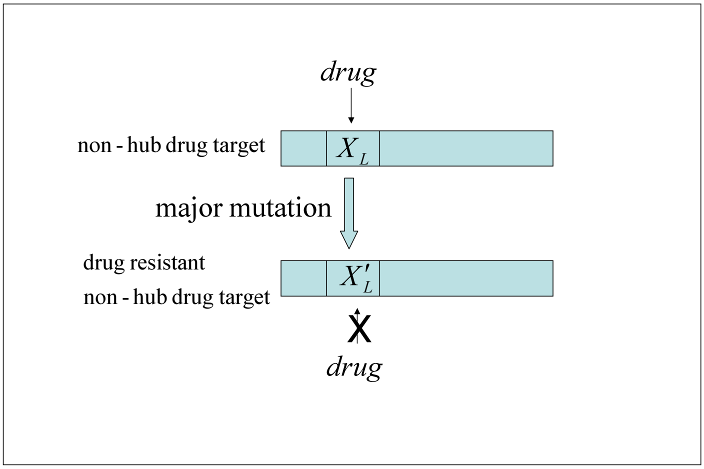


Fig. 1. Non-hub drug target

4.2 Hub Drug Target

This subsection treats with a hub drug target. Let intracellular agent M_i $(1 \leq i \leq k)$ to bind amino-acid Y_i at a position i on the amino-acid sequence of the drug target.

In order for a hub drug target molecule to develop resistance to a drug, it requires that a major mutation $(X_L \rightarrow X'_L)$ arises at a drug binding position L, and any mutations at intracellular agent-binding positions i $(1 \leq i \leq k)$ must not arise (Fig. 2). Under such a condition, a drug resistance probability of the hub drug target, P_h is expressed as bellow,

$$P_h = \sum_{X'_L(\neq X_L)} P(X_L \rightarrow X'_L) \prod_{i=1}^{k} P(Y_i \rightarrow Y_i). \tag{2}$$

5 Discussions

This section demonstrates that the hub drug target has stronger resistivity than the non-hub drug target on the development of the drug resistance.

From the comparison of the Eq.(1) with Eq.(2), as $P(Y_i \rightarrow Y_i)$ is less than one,

$$P_{nh} > P_h \tag{3}$$

succeeds.

Interestingly, the probability P_h decreases exponentially with increase of k representing the degree of hub-ness. That means that hub-ness of the drug target is one of remarkable factors to reduce the risk of the drug resistance.

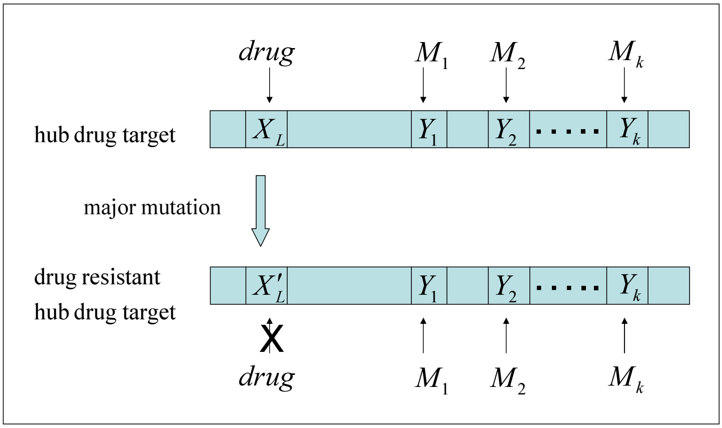


Fig. 2. Hub drug target

6 Conclusions

This study has focused on the hub-ness of drug targets and has shown the drug targets are classified into the two types: non-hub type and hub one by referring to a case of HIV-1 drug targets. In consideration with the classification, it has developed the expression for the probability of these classified drug targets developing drug resistance, and it has demonstrated a hub drug target has stronger resistibility on the development of the drug resistance than a non-hub one.

References

1. Volberding, P.A., Sande, M.A., Lange, J., Greene, W.C.: Global HIV/AIDS Medicine. Elsevier Inc., Amsterdam (2008)
2. Chiryo no tebiki 12th edn., http://www.hivjp.org/
3. Sugiura, W.: Progress in antiretroviral drugs. Virus 55, 85–94 (2005)
4. Harada, K., Ishida, Y.: A hub gene in a HIV-1 gene regulatory network is a promising target for anti-HIV-1 drugs. In: Proc. of AROB 14th, pp. 522–525 (2009)
5. Ptak, R.G., et al.: Cataloging the HIV Type 1 Human Protein Interaction Network. AIDS Research And Human Retroviruses 24(12), 1497–1502 (2008)
6. Zambrowicz, B.P., Sands, A.T.: Modeling drug action in the mouse with knockouts and RNA interference. Drug Discov. Today Targets 3, 198–207 (2004)
7. Winzeler, E.A., et al.: Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. Science 285, 901–906 (1999)
8. Giaever, G., et al.: Functional profiling of the Saccharomyces cerevisiae genome. Nature 418, 387–391 (2002)
9. Baraba'si, A.L., Oltvai, Z.N.: Network biology: understanding the cell's functional organization. Nat. Rev. Genet. 5, 101–113 (2004)
10. Albert, R., Jeong, H., Barabasi, A.L.: Error and attack tolerance of complex networks. Nature 406, 378–382 (2000)

11. Ooi, S.L., et al.: Global synthetic-lethality analysis and yeast functional profiling. Trends Genet. 22, 56–63 (2006)
12. Denome, S.A., Elf, P.K., Henderson, T.A., Nelson, D.E., Young, K.D.: Escherichia coli mutants lacking all possible combinations of eight penicillin binding proteins: viability, characteristics, and implications for peptidoglycan synthesis. J. Bacteriol. 181, 3981–3993 (1999)
13. Janoir, C., Zeller, V., Kitzis, M.D., Moreau, N.J., Gutmann, L.: High-level fluoroquinolone resistance in Streptococcus pneumoniae requires mutations in parC and gyrA. Antimicrob. Agents Chemother. 40, 2760–2764 (1996)
14. Hopkins, A.L.: Network pharmacology: the next paradigm in drug discovery. Nat. Chem. Bio. 4(11), 682–690 (2008)
15. Dancey, J.E., Chen, H.X.: Strategies for optimizing combinations of molecularly targeted anticancer agents. Nat. Rev. Drug Discov. 5, 649–659 (2006)
16. Han, J.D., et al.: Evidence for dynamically organized modularity in the yeast proteinprotein interaction network. Nature 430, 88–93 (2004)
17. Joy, M.P., Brock, A., Ingber, D.E., Huang, S.: High-betweenness proteins in the yeast protein interaction network. J. Biomed. Biotechnol. 2, 96–103 (2005)
18. Yu, H., Kim, P.M., Sprecher, E., Trifonov, V., Gerstein, M.: The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. PLoS Comput. Biol. 3, e59 (2007)
19. Hwang, W C., Zhang, A., Ramanathan, M.: Identification of information flowmodulating drug targets: a novel bridging paradigm for drug discovery. Clin. Pharmacol. Ther. (published online) doi:10.1038/clpt.2008.12

Asymmetric Phenomena of Segregation and Integration in Biological Systems: A Matching Automaton

Yoshiteru Ishida and Tatsuya Hayashi

Department of Knowledge-Based Information Engineering,
Toyohashi University of Technology
Tempaku, Toyohashi 441-8580, Japan
http://www.sys.tutkie.tut.ac.jp

Abstract. Seemingly conflicting phenomena of segregation and integration have been observed both in the immune system and the neural system, and possible mechanisms have been studied relating to learning and adaptation. Inspired by the *Stable Marriage Problem* whose solutions may exhibit both segregation and integration among agents, we propose a working model of matching automata which take order of preference as inputs and the resultant matching among agents as outputs. In this tentative model, we try to simulate the integration taking the switching experienced in the *Necker Cube* as an example, while the segregation is built into the restrictions of the model.

Keywords: neural system, immune system, segregation and integration, stable marriage problem, matching automaton, *Necker Cube*.

1 Introduction

Integration and segregation, although apparently conflicting, also seem to be compatible, for they are to be found in many biological processes. The immune system is known to adapt to the antigenic environment: the system will increase the intensity of reaction to frequent challenges (integration) and yet will not lose the reaction to other challenges (segregation).

The neural system has been extensively studied from the viewpoints of integration/segregation [1, 2], which motivated this work. As has been pointed out [1, 2], integration/segregation is observed in the visual area between the sensory sheet and recognition part for higher processing, where the sensory sheet consists of neurons that respond to specific orientations of stimuli. Integration/segregation can also be found in the audio area, with the sensory sheet consisting of neurons responding to specific frequencies.

While integration/segregation has been the focus of studies of biological systems, the *Stable Marriage Problem* (*SMP*) [3, 4, 5] has been studied extensively in fields such as discrete mathematics, algorithms, and economics. Because group aggregation and decomposition play an important role not only in solving the *SMP* but also in observing solutions, we propose a model of matching automata that includes the *SMP* framework, with the aim of shedding new light on the *SMP* and biological modeling.

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 789–796, 2009.

Section 2 illustrates the definition of the *Stable Marriage Problem*. Based on the *SMP*, section 3 presents the basic model of matching automata. Section 4 discusses the implications of the model focusing on the segregation and integration process found in the neural system and the immune system.

2 Background: Stable Marriage Problem

The *Stable Marriage Problem* (*SMP*) assumes *n* women and *n* men, each of whom has an ordered preference list (or a ranking) without tie to the opposite sex. As in the example shown in Fig. 1, the man m_2 has a ranking (3, 2, 1, 4), which means that m_2 likes w_3 best, and he prefers w_3 to w_2, w_2 to w_1, and w_1 to w_4. One could say that there is an *injection* (one to one, but not necessarily onto) mapping from a set of women (men) to an element of a permutation group of size *n* such as shown in the ranking by each person (Fig. 1).

Under the above assumptions, the *SMP* seeks complete matching between *n* women and *n* men (a *bijection* from *n* women to *n* men), which satisfies *stability*. The stability requires the concept of *blocking pairs*. Two pairs (m_i, w_p) and (m_j, w_q) are blocked by the pair (m_i, w_q) if m_i prefers w_q to w_p and w_q prefers m_i to m_j. A complete matching without being blocked is called *stable* matching.

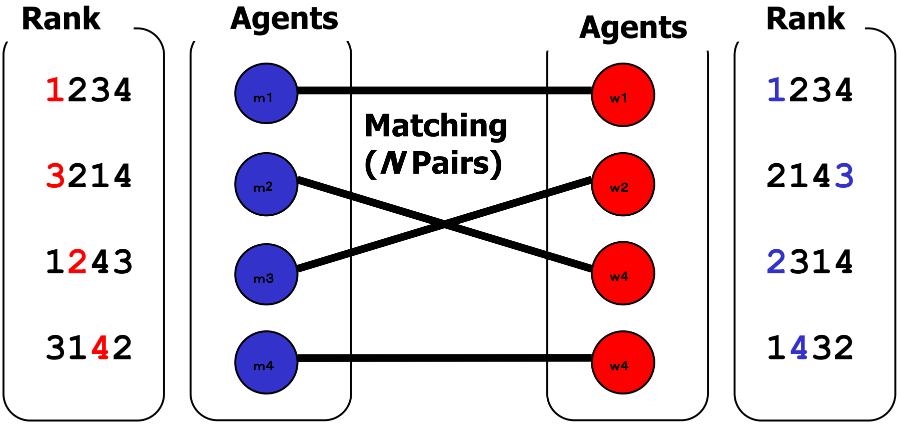


Fig. 1. An example of a bipartite graph indicating a matching of *SMP* with size 4. Agents on the right are women, and those on the left are men. Each agent has its own ranking of preference for the members of the opposite sex.

2.1 A Network of Matching Solutions

Although instances of men's preference and women's preference can be expressed by networks, we choose to express matching solutions on networks. In the network, each node expresses matching and an edge between two nodes indicates that the matching corresponding to the node can be realized by exchanging partners in two pairs of another linked node (Fig. 2). With two sets of n men and n women above, let us consider the following two matchings M_1 and M_2:

$$M_1 = \{(m_1,w_1),(m_2,w_4),(m_3,w_3),(m_4,w_2)\}$$
$$M_2 = \{(m_1,w_1),(m_2,w_3),(m_3,w_4),(m_4,w_2)\}$$

The matching M_2 can be attained by exchanging the partners in two pairs: (m_2,w_4), and (m_3,w_3) in the matching M_1, thus two nodes corresponding to these two matchings are linked in the network. We will call the network *complete* when the network includes all possible matchings as nodes and all possible partner-exchanges as links.

2.2 Coordinates for Network Visualization

The motivation for visualizing matching solutions as a network is to make it easier to perceive the regularities and symmetries, which are unseen otherwise. The selection of appropriate coordinates (and its scale) is crucial for this purpose. Here, we will use simple and natural coordinates. For example, men's satisfaction P_m is defined as follows:

$$P_m = \sum (n+1-R_{mi})\ ,\ P_w = \sum (n+1-R_{wp})$$

where n is the size of the *SMP* and R_{mi} is the man m_i's rank (an integer ranging from 1 to n where 1 means the most favorite) to the current partner in the matching M. Women's satisfaction P_w is similarly defined with R_{wp} being the woman w_p's rank to the partner in the matching.

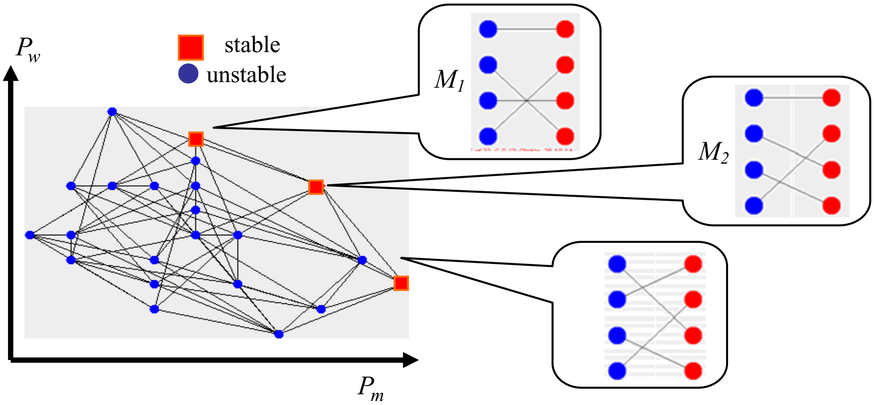


Fig. 2. Network visualization of an instance of SMP with size 4. Two coordinates of men's satisfaction P_m and women's satisfaction P_w are used.

3 Basic Model: A Matching Automaton

The structure of solutions (matchings) including stable ones and the process of solving the *SMP* exhibit similar phenomena to those observed in biological processes.

Among them, we focus on the seemingly conflicting phenomena of integration and segregation.

In order to keep the model minimal, we will reserve all of the assumptions and components in the *SMP* intact. That is, the matching automaton consists of two parts of *N* agents each of which has a preference list (without tie) to the members of the other parts. The preference is considered to be the input to the automata, and the resulting matching is used as the output. In forming the matching, *complete matching* must be obtained. That is, neither *singlehood* nor p*olygamy* is allowed.

Thus, segregation is already built into the restriction that the output must satisfy. Hence, hereafter, we demonstrate that the matching automata (equivalent to *SMP*-solving automata) with the segregation built into the definition can exhibit integration when the input to the automata is properly devised. The automaton also uses optimality for one set of agents (*women-optimal* or *men-optimal*) to switch the output configurations.

As an illustration, let us explain the simplest matching automaton of a switching gate. The gate consists of two by two agents (*SMP* with size 2) and each agent in a set has a distinct preference to avoid symmetry; furthermore, each pair of agents has no first rank assignment (no *mutual infatuation*) to avoid fixation of the pair in matching. These restrictions lead to the following preference structure (expressed by a preference matrix { a_{ij} } where the element in the $i^{\text{-th}}$ row and $j^{\text{-th}}$ column a_{ij} is defined to be $R_{mi,wj}$ / $R_{wj,mi}$ and $R_{mi,wj}$ to be a rank of m_i to w_j):

2/1, 1/2
1/2, 2/1.

With this preference as input, the behavior of the matching automaton is grasped by the matching network (Fig. 3). Note that only two matchings exist and both of them are stable. When the automaton seeks the women-optimal solution, then the matching will be the one on the upper left. On the other hand, if the automaton seeks the men-optimal solution, the matching will be switched to the one on the lower right.

The switching gate with two channels may be devised by adding other agents (for each part) as auxiliary agents to control the switch. Then, switching will be done by changing the input (preference) including these controlling agents. The technical

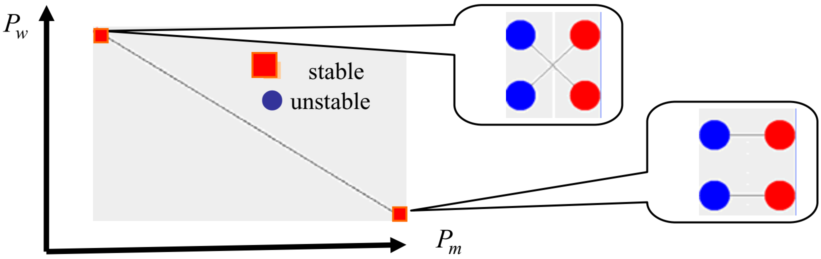


Fig. 3. An example of a switching gate realized by the matching automaton of 2 by 2 agents. Two squares indicate stable matchings. The matchings will be switched among one another by changing the mode of automaton, i.e., women-optimal or men-optimal.

details of devising a switching gate implementation and gates other than switching is out of the scope of this paper and will be discussed elsewhere.

3.1 Matching Automata with Specific Preferences

Although the preferences can be any permutation out of a list of $n!$, we use an ordered preference: the cyclic preference which can be generated by shifting one digit in the ordered list of 1, 2, 3, ..., n where n is the number of agents for one part.

If agents in one part use the cyclic preference, the satisfaction level of agents belonging to that part will be divided into n distinct levels. In the following example, since both sensory and recognition parts adopt the cyclic preference, the level of satisfaction is divided into six for agents in both parts.

It should be noted that the matching network on the coordinate will be symmetric for these two parts if they adopt the same cyclic preference. Further, if both parts adopt the cyclic preference, then it can easily trigger a *domino effect*. That is, replacing a partner would lead to another replacement and so on until all the agents replace their partners. Because of the domino effect, the switching mechanism implemented by the above matching automaton can have two stable matchings. Thus, without changing the modes (*women-optimization* and *men-optimization*), two stable matchings can be switched by forcing an agent to replace the partner, which would in turn lead to another replacement, hence leading to a chain of replacements.

3.2 An Illustrative Example of *Necker Cube*

The matching automaton that will explain the switching experienced when observing the *Necker Cube* (e.g. [6]) will be devised by extending the previous switching gate. It is straightforward to implement a *k by k* switching gate, and there are many different ways of doing so with the automaton. We use a 6 by 6 gate here.

Although the cube consists of 12 line segments, we focus on six of them (we could alternatively use six faces). These six segments are numbered as shown in Fig. 4. These numbers also indicate the numbers (labels) of the agents in a sensory part (left-side agents in the bipartite graph shown in Fig. 4). When the agent i in the sensory part is paired to the agent j f (b) in the recognition part (right-side in the bipartite graph shown in Fig. 4), then the automaton understands that the bar i is in the front (behind). Thus, for the automaton to output two understandings in the assignment shown in Fig. 4, they form the two matchings as shown in Fig. 4.

As an example of the matching automaton realizing the above switching, the preference matrix is as follows:

1/2, 2/1, 3/6, 4/5, 5/4, 6/3
6/3, 1/2, 2/1, 3/6, 4/5, 5/4
5/4, 6/3, 1/2, 2/1, 3/6, 4/5
4/5, 5/4, 6/3, 1/2, 2/1, 3/6
3/6, 4/5, 5/4, 6/3, 1/2, 2/1,
2/1, 3/6, 4/5, 5/4, 6/3, 1/2.

The implementation of the 6 by 6 switching gate uses a cyclic preference for both sensory and recognition parts.

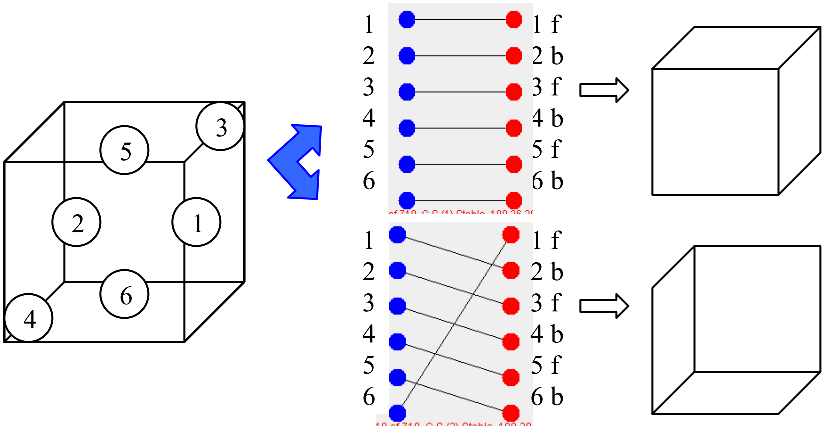


Fig. 4. Perceptual switching experienced when observing the *Necker Cube*. A cube whose six line segments are numbered (left). Two matchings showing two possible interpretations (middle). Two cubes corresponding to the matchings (right).

The behavior of the automaton with the above preference matrix is illustrated by the matching network shown in Fig. 5. It is known that two stable matchings satisfy the requirements specified in Fig. 4. Again, these two matchings can be switched by switching between the optimality of agents in the sensory part (men-optimality) and the optimality of agents in the recognition part (women-optimality).

Since sensory and recognition parts adopt the cyclic one, the levels of satisfaction are divided into six for both coordinates (Fig. 5, left). Also, it can be observed that the matching network on the coordinate is symmetric for these two parts (Fig. 5, left).

With the switching among multiple stable solutions, not only the switching in the *Necker Cube* but many other gestalt phenomena may be explained similarly. Interestingly, simultaneous switching, which could be explained by integration of similar agents in a sensory part, has been attained without direct connection among these agents in the sensory part. This was achieved by local, selfish and autonomous acts of agents.

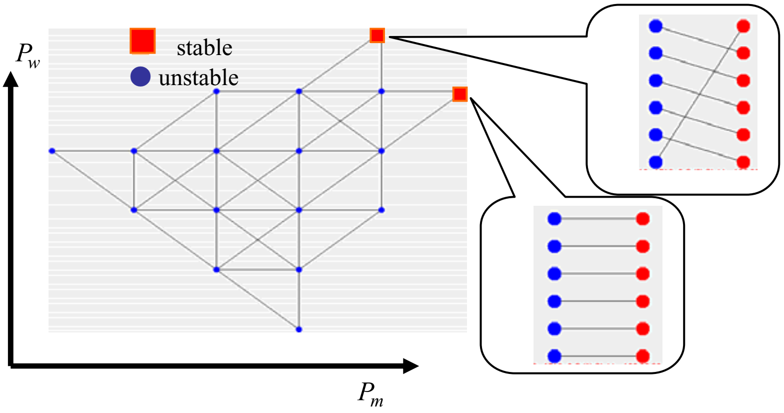


Fig. 5. A realization of matching automata that would explain the switching phenomenon experienced when observing the *Necker Cube*

4 Discussion

Since this note focuses on the simplest model, we try to explain segregation/integration phenomena without violating the assumptions and concepts defined in the *Stable Marriage Problem*. It is demonstrated that even with this simplest model, the segregation/integration phenomena have been analogously explained. Other than the phenomena, the simplest model further has the potential for simulating not only phenomena related to switching in recognition such as optical illusions experienced in observing the *Necker Cube* and *trompe l'oeil*, but also those related to substitution in functional parts due to plasticity (plasticization) such as a sensory substitution experienced when one sense is damaged.

Although the current simplest model of matching automata seems to have sufficient potential to explain broad phenomena found in recognition, memory, learning, and adaptation, it is tempting to involve signal forwarding (not only receiving the signal as a sensory unit, it would forward a similar or complementary one to other units) and connection to investigate how the scope of the model may be expanded or the qualitative differences that the model would gain.

When building more detailed models specific to the immune system or neural system, the model may require elaboration such as direct linking among agents in the same part (for the neural system) and agent reproduction and diversification (for the immune system). However, we can use many results in the *SMP* by keeping its framework intact. Violations of assumptions such as the same number of agents for both parts can be easily avoided by including dummy agents. Furthermore, there have been many extensions, generalizations and modifications for the *SMP*. The number of parts, for example, can be single (Stable Roommate Problem) or three (Three-Party version).

5 Conclusion

Integration and segregation, although appearing to conflict with each other at first glance, can be compatible at least through the model of this paper. It remains unknown, however, whether these two phenomena may be accompanying or not, since the integration we observed in the model can always be stopped from occurring by randomizing the preference, and the segregation is built into the model beforehand.

It is rather surprising that the simplest model of matching automata (which keeps the framework of the *SMP* intact) can simulate the seemingly conflicting phenomena of integration and segregation. Even the switching as experienced in the *Necker Cube* has been explained. However, the simulation is rather superficial (preferences are arbitrary and artificial) and furthermore, segregation is already built into the definition of the automaton. True challenges for matching automata will be to modify the model involving details to be more specific to the target system as well to expand the scope of relations with other phenomena while maintaining the simplicity of the model.

Acknowledgments. I am grateful to Yoshihisa Morizumi who programmed a network visualization tool for the *Stable Marriage Problem*. Without it, not only the figures of this paper but also the construction of matching automata would not have been possible.

References

1. Tononi, G., Sporns, O., Edelman, G.M.: A complexity measure for selective matching of signals by the brain. Proc. Natl. Acad. Sci. USA 93, 3422–3427 (1996)
2. Sporns, O., Chialvo, D., Kaiser, M., Hilgetag, C.C.: Organization, development and function of complex brain networks. Trends in Cognitive Sciences 8, 418–425 (2004)
3. Gale, D., Shapley, L.S.: College admissions and the stability of marriage. American Mathematical Monthly 69, 9–15 (1962)
4. Gusfield, D., Irving, R.W.: The Stable Marriage Problem: Structure and Algorithm. MIT Press, London (1989)
5. Knuth, D.E.: Stable marriage and its relation to other combinatorial problems: An introduction to the mathematical analysis of algorithms. CRM Proceedings & Lecture Notes 10 (1997)
6. Marr, D.: Vision: A Computational Investigation into the Human Representation and Processing of Visual Information. W.H. Freeman, New York (1983)

Adaptive Forecasting of High-Energy Electron Flux at Geostationary Orbit Using ADALINE Neural Network

Masahiro Tokumitsu[1], Yoshiteru Ishida[1], Shinichi Watari[2], and Kentarou Kitamura[3]

[1] Department of Electronic and Information Engineering, Toyohashi University of Technology,
1-1 Hibarigaoka, Tempaku, Toyohashi, Aichi, 441-8750 Japan
[2] Applied Electromagnetic Research Center, National Institute of Information and
Communications Technology, 4-2-1 Nukuikitamachi, Koganei, Tokyo, 184-8795 Japan
[3] Department of Mechanical and Electrical Engineering, Tokuyama College of Technology,
Gakuendai, Shuna, Yamaguchi, 745-8585 Japan

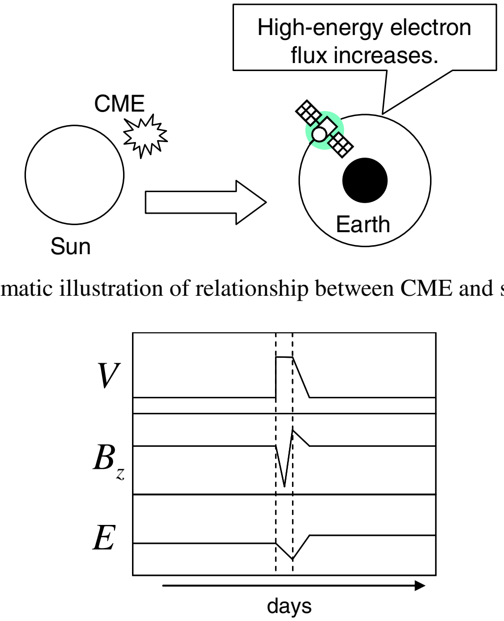
Abstract. High-energy electron flux increases in the recovery phase after the space weather events such as a coronal mass ejection. High-energy electrons can penetrate circuits deeply and the penetration could lead to deep dielectric charging. The forecast of high-energy electron flux is vital in providing warning information for spacecraft operations. We investigate an adaptive predictor based on ADALINE neural network. The predictor can forecast the trend of the daily variations in high-energy electrons. The predictor was trained with the dataset of ten years from 1998 to 2008. We obtained the prediction efficiency approximately 0.6 each year except the first learning year 1998. Furthermore, the predictor can adapt to the changes for the satellite's location. Our model succeeded in forecasting the high-energy electron flux 24 hours ahead.

Keywords: Adaptive Learning, Neural Network, High-energy Electron, Dielectric Charging of Spacecraft, Space Weather.

1 Introduction

Satellites are important social infrastructures. There are high-energy electrons at Geostationary Earth Orbit (GEO). High-energy electrons could take charge to the surface of cables and circuits in the spacecrafts. Then, high-energy electrons can penetrate circuits deeply and the penetration could lead to deep dielectric charging. For example, the Intelsat K spacecraft at GEO lost altitude control due to the failure of the momentum wheel control circuitry on January 20, 1994. According to the observations by Geostationary Operational Environmental Satellites (GOES), high-energy electron flux largely enhanced during the Intelsat anomalies. The analysis of specialists revealed that these spacecraft anomalies occurred due to dielectric charging by the high-intensity and long-duration enhancement of high-energy electrons [1, 2]. These studies also reported that the spacecrafts anomalies at GEO are associated with enhancement in high-energy electron flux.

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 797–804, 2009.



(a) Schematic illustration of relationship between CME and spacecraft.

(b) Schematic plot of parameters correlated with CME. The ***V***, ***Bz*** and ***E*** are corresponded to the solar wind speed, the south-north component of interplanetary magnetic field and the high-energy electron flux.

Fig. 1. Schematic illustration and plot on CME

The enhancement of high-energy electron flux is known to be correlated with the solar activities such as Coronal Mass Ejection (CME) and the coronal hole on the surface of the sun. The electron flux varies in two phases: main phase and recovery one (Fig.1). During the main phase, the electron flux rapidly decreases; and after the phase, it increases significantly. The problem is that the high flux level causes the irreparable damage to the instruments on satellites in the recovery phase of geomagnetic storms.

The dynamics of high-energy electrons is under investigation [3], although the enhancement has been observed by the specialists in space physics. Many studies, however, have reported the enhancement of high-energy electron is correlated with high-speed solar wind [4, 6]. The high-energy electron flux is known to be controlled by the solar wind. Furthermore, the north-south component of the interplanetary magnetic field (IMF) is also known to be another important parameter that promotes flux enhancement.

Many predictors for high-energy electron flux at GEO have been proposed. The motivation for developing the predictor is to protect spacecrafts from the deep dielectric charging. The linear prediction filter with a statistical approach has been developed [5]. The model is capable of predicting the daily averages of electron flux at GEO. Koons and Goorney have also investigated a predictor of the daily averaged flux at GEO using artificial neural networks [7]. An advanced predictor similar to the one has been developed by the recurrent neural network [9, 10]. These researches are supported by various techniques and algorithms. The advanced algorithms and techniques allow us the forecasting of the high-energy electron flux.

Adaptive predictors to the space environment are needed, since the solar wind speed and high-energy electron flux will change due to the solar activities. The predictors of the earlier studies are not capable of adapting to the changing environment. The predictors require a training (learning the new environment) to deal with the changes. Furthermore, the huge data are required in the predictors.

This paper proposes an adaptive predictor using the ADALINE Neural Network (ANN) [11]. The neural network is capable of obtaining the input/output map by training. The ANN can learn the variation of the observed data. Because of its simplicity, the ADALINE has been applied to many applications; SIR estimation system [8], for example. Only a little train is needed, since the model only uses the data of the difference with one hour averaged data. The model can predict the electron flux allowing an adaptation to the changes in the space environment.

2 Basic Model

2.1 Training Data and Forecast Data

We use one hour averaged data of the solar-wind and the high-energy electron flux at GEO. The solar wind data observed by Advanced Composition Explorer (ACE) satellite are obtained from the OMNI-2 database [13] in the National Space Science Data Center (NSSDC), the National Aeronautics and Space Administration/Goddard Space Flight Center.

The electron flux data observed by the GOES satellite are obtained from the National Geophysical Data Center (NGDC), and the National Oceanic and Atmospheric Administration (NOAA). We obtain both data during the period from January 1, 1998 to December 31, 2008, thus 10 years in total. We use the data to train the predictor.

2.2 Predictor with ADALINE Neural Network

The ADALINE neural network (ANN) is an artificial neural network [11]. The network structure of the ANN consists of input neurons and output neurons with a linear transfer function. We use the simplest one to predict the high-energy electron flux (Fig.2). In Fig.2, let the I_t and E_{24} denote respectively the t th input data vector and the high-energy electron flux 24 hours ahead.

In the neural networks, choosing appropriate input parameters is critical for the high-precision prediction result. We need to select the input parameters correlated with high-energy electrons. We choose the four input parameters: solar wind speed V , a north-south component B_z of IMF, the current high-energy electron flux E and the universal time UT . Thus, the t th input parameters are expressed as an input vector $I_t = (E, V, B_z, UT)$.

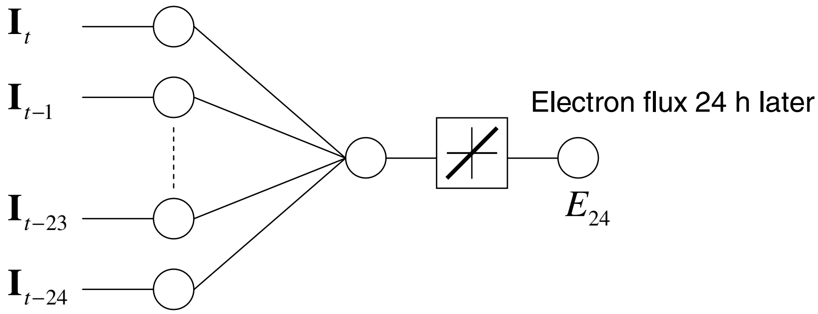


Fig. 2. Configuration of ADALINE neural network predictor. Predictor computes high-energy electron flux 24 hours ahead by the input data using past 24 hours data.

The current high-energy electron flux is chosen to reflect the flux level to the forecast. We add UT as an input parameter to involve a large diurnal variation in the electron flux, since the flux level differs according to location at GEO. The growth in the high-energy electron flux is strongly correlated with the solar wind speed; hence, the solar wind speed is selected as the input parameter. The north-south component of IMF is chosen, because of the value gradually varies to the north and south when the CME happens or the coronal hole appears on the surface of the sun.

The predictor is required to support the time series data, because the electron flux is affected by the variation in the past. This requirement can satisfy with the ANN. We arranged the input neurons with amount of 24 hours past and current. The predictor computes high-energy electron flux based on the past 24 hours and the current flux.

2.3 Validation Method

We need to measure the performance of the predictor. The performance is evaluated by comparing the predictions with the observations. The performance is tested by the three indices: Correlation Coefficient (CC), Mean-Square Error (MSE) and Prediction Efficiency (PE). Let x_i and f_i represent respectively the forecasted and observed data. The N is total count of the data. The PE and MSE are expressed as:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(f_i - x_i\right)^2 \text{, and} \tag{1}$$

$$PE = 1 - \frac{MSE}{VAR}, \tag{2}$$

$$\text{where } VAR = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2 \text{, and} \tag{3}$$

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \,. \tag{4}$$

3 Simulations and Results

3.1 Data Handling

The data observed by the satellites could be missing data due to the instruments troubles by the space weather events and/or various reasons for the operations. We regard the data as missing where the interval of the missing exceeds two hours. The missing data are interpolated if the observation down time is less than three hours. We exclude the missing data in training and simulations.

3.2 Simulation Method

We construct the predictor for 24 hours ahead forecast of the high-energy electron flux at GEO. The advance warnings are useful as space weather monitors, especially to satellite and communications satellite. We train the predictor with the learning rate of 0.005. A linear transfer function is used as the transfer function in the output layer.

We train the predictor by setting to the electron flux 24 hours ahead as the target output. The total count of the data available to the predictor is 84563, whereas the total count for the simulation data is 84539. Both dataset exclude the missing data and have already interpolated with the spline function. All of the data are arranged with date order. We input the data to the predictor sequentially from 1998 to 2008.

Our training and simulation of the predictor consists of two phases. The predictor learns and forecasts alternatively. The first phase is used for the training of the network. The predictor is given past 24 hours data as input data and the current high-energy electron flux as the target data. In the second phase, the network predicts the high-energy electron flux 24 hours ahead by the current and past 24 hours data. The simulation results are validated above method (Sec.2.3).

The forecast of the alert level of the high-energy electron flux is meaningful to protect the spacecrafts. The spacecrafts will become anomalies, when the high-energy electron flux enhances to the alert level. Thus, we estimate the prediction results of the *MSE* by dividing into two levels: $E_{24} >= 4$ and $E_{24} <= 2$. We need to clarify the errors between the predictions and observations data. The MSE_H and MSE_L are respectively corresponded to the flux of low and high level.

3.3 Results

Figure 3 shows the forecast performance form 1998 to 2008. In 1998, the *PE* was the worst during the forecast period. Because of the year was first for the training of the predictor. In this year, the predictor adjusted its weight to forecast the reliable output. After that, the predictor improved the *PE* to approximately 0.6. Our model could adapt to the input and target data by training. Then, the model succeeded in forecasting the high-energy electron flux 24 hours ahead.

Furthermore, the predictor could keep the *CC* approximately 0.8 each year expect in 1998. The forecast results were positively correlated with the observed data. Thus, the prediction of the daily variations of the high-energy electron flux was successful.

The MSE_L of the low flux largely increased in 2003 and 2005. During the main phase of the geomagnetic disturbances, forecasting of the high-energy electron flux is difficult since the flux rapidly decreases. Our predictor could not involve the rapid decreases in the flux by disturbed solar winds during the 24 hours interval. Thus, the geomagnetic disturbances by the CME and the coronal hole caused the results since those kinds of the space weather events occurred more times than usual.

At the end of 2006, the GOES-10 was transferred to the new location at 60° west. The predictor has needed to retrain because the environment of surrounding the GOES-10 changed. The local time of the satellite also was needed to consider because its location differs from one before transferring. However, the *PE* values did not indicate the significant change after the satellite moved. The predictor could adapt to the new environment.

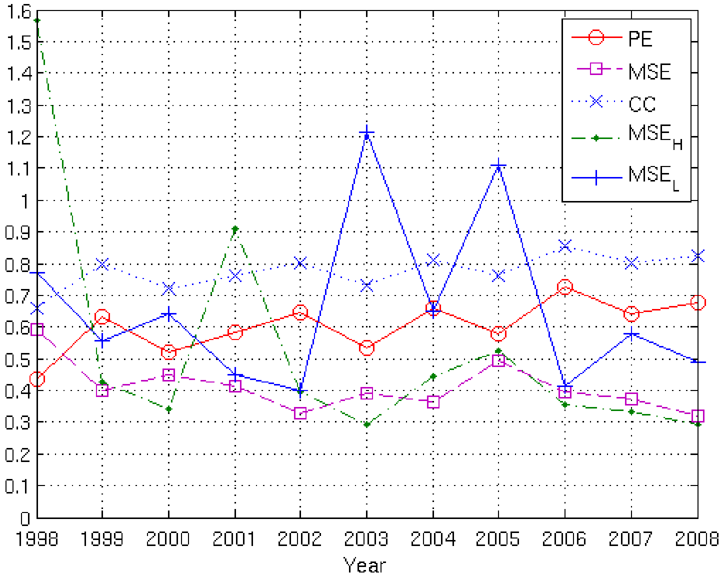


Fig. 3. Forecast performance from 1998 to 2008

4 Discussion

We investigated the predictor for the forecast of the high-energy electron flux at GEO. The predictor is constructed with the ADALINE neural network. The predictor could forecast of the daily variations of the high-energy electron flux. Furthermore, the predictor could also forecast the high flux in the recovery phase of geomagnetic storms. However, the model has only a little ability to forecast the low flux since the predictor does not involve the rapid decreases during the 24 hours interval.

The major advantage of the model is an adaptation to the new environments for the satellites. The dataset for forecasting covers mostly one solar cycle period (The solar cycle period is 11 years). The predictor could adapt to the changes of the high-energy electron flux by the solar activity. Furthermore, our model can involve the changes of the flux when the satellite moves to the new environment. In Fig. 1, the predictor could perform the forecast of the flux in 2006 after the satellite moved to the new location. Then, the predictor can adapt to the new environment and forecast the flux without degrading its performance.

The earlier studies [9, 10] do not provide the adaptation to the new data for the predictor. Actually, the satellites are transferred to the new location by various operational reasons. After transferring, the predictors are required to learn the changes of the new environments. The model allows us to solve the relocation problem of the satellites for the forecast since the predictor adapt to the changes. Therefore, the model can apply to the satellites at GEO if they are transferred in the future. However, the learning rate for the training should be chosen carefully because the performance depends on the learning rate.

Sometimes, the observed data contains the irrelevant data due to the failures of sensors on satellites. The model cannot forecast the high-energy electron flux if any sensors behaves as the faulty one. Thus, the detection of the sensor's failure is also a crucial problem to do the exact prediction. The predictor is required to involve the robustness and adaptation to the failures of the sensors. In the related study [12], the Dynamical Relational Networks (DRN) has been proposed. The DRN technique can diagnose the sensors faulty or non-faulty with autonomous distributed approach. The study [12] demonstrated the online diagnosis systems of combustion systems in automobile engines. Likewise, we should consider that the predictor for space weather also needs to involve the robustness and the adaptation for the failures of the sensors.

5 Conclusion

We have investigated the adaptive predictor for the high-energy electron flux at GEO. The ADALINE neural networks are used in the implementation. The predictor has been trained by the adaptive learning algorithm. In the simulation results, our model obtained the *PE* and the *CC* approximately 0.6 and 0.8 respectively, except the training phase in 1998. Furthermore, the predictor could adapt to the changes of the environment after the GOES-10 was transferred to the new location. Thus, our predictor successfully reflected the changes of the data for the forecast.

Acknowledgements

This work was supported by The Global COE Program "Frontiers of Intelligent Sensing", from the ministry of Education, Culture, Sports, Science and Technology. The OMNI data were obtained from the GSFC/SPDF OMNIWeb [13] interface at http://omniweb.gsfc.nasa.gov/. The GOES-10 data were obtained from the SPIDR by NOAA/NGDC, the interface at http://spidr.ngdc.noaa.gov/spidr/.

References

1. Baker, D.N., Kanekal, S., Blake, J.B., Klecker, B., Rostoker, G.: Satellite anomalies linked to electron increase in the magnetosphere. EOS Transactions 75, 401–405 (1994)
2. Baker, D.N.: Satellite Anomalies due to Space Storms. In: Daglis, I.A. (ed.) Space Storms and Space Weather Hazards, pp. 285–311 (2001)
3. Liemohn, M.W., Chan, A.A.: Unraveling the Causes of Radiation Belt Enhancements. EOS Transactions 88, 425–426 (2007)
4. Kataoka, R., Miyoshi, Y.: Flux Enhancement of Radiation Belt Electrons during Geomagnetic Storms Driven by Coronal Mass Ejections and Corotating Interaction Regions. Space Weather 4, S09004 (2004)
5. Baker, D.N., McPherron, R.L., Cayton, T.E., Klebesadel, R.W.: Linear Prediction Filter Analysis of Relativistic Electron Properties at 6.6 Re. J. Geophys. Res. 95(A9), 15133–15140
6. Baker, D.N., Klebesadel, R.W., Higbie, P.R., Blake, J.B.: Highly Relativistic Electrons in the Earth's Outer Magnetosphere. I - Lifetimes and Temporal History 1979-1984. J. Geophys. Res. 91, 4265–4276 (1986)
7. Koons, H.C., Gorney, D.J.: A Neural Network Model of the Relativistic Electron Flux at Geosynchronous Orbit. J. Geophys. Res. 96, 5549–5556 (1991)
8. Ardalani, N., Khoogar, A., Roohi, H.: A Comparison of ADALINE and MLP Neural Network-based Predictors in SIR Estimation in Mobile DS/CDMA Systems. Computing and Technology 9, 145–150 (2005)
9. Fukata, M., Taguchi, S., Okuzawa, T., Obara, T.: Neural network prediction of relativistic electrons at geosynchronous orbit during the storm recovery phase: effects of recurring substorms. Annales Geophysicae 20, 947–951 (2002)
10. Watari, S., Tokumitsu, M., Kitamura, K., Ishida, Y.: Forecast of High-energy Electron Flux at Geostationary Orbit Using Neural Network. In: The 26th International Symposium on Space Technology and Science, ISTS 2008 (2008)
11. Widrow, B., Lehr, M.A.: Perceptrons, adalines, and backpropagation. The handbook of brain theory and neural networks, 719–724 (1998)
12. Ishida, Y.: Designing an Immunity-Based Sensor Network for Sensor-Based Diagnosis of Automobile Engines. In: Gabrys, B., Howlett, R.J., Jain, L.C. (eds.) KES 2006. LNCS (LNAI), vol. 4252, pp. 146–153. Springer, Heidelberg (2006)
13. King, J.H., Papitashvili, N.E.: Solar wind spatial scales in and comparisons of hourly Wind and ACE plasma and magnetic field data. J. Geophys. Res (Space Physics) 110(A9), A02104 (2005)

A Note on Biological Closure and Openness: A System Reliability View

Yoshiteru Ishida

Department of Knowledge-Based Information Engineering,
Toyohashi University of Technology
Tempaku, Toyohashi 441-8580, Japan
http://www.sys.tutkie.tut.ac.jp

Abstract. Inspired by metabolic closure and its mathematical realization as a fixed point of *f(f) = f* where *f* is an operator, operand, and result, we pursue the possibility of reproduction closure of organisms. We seek an information aspect of reproduction closure, expecting an organizing principle of information (entropy) in living organisms. To remain reliable as a system with unreliable components, living organisms use reproduction involving the description (genotype). A reliability view of self-reproduction with a description will be compared to von Neumann's complexity decrease principle in building automata. Asymmetry of complexity decrease would indicate that the self-reproduction in his model is not reversible, hence suggesting entropy generation (negative entropy leak). Although errors in copying description would lead to threats of cancer, allowance of a certain level of error would lead to possible adaptation recognizing the openness of biological systems.

Keywords: metabolic closure, biological closure, system reliability, entropy, self-reproduction, quasi-species, cancer, apoptosis, immune system.

1 Introduction

This note is motivated by Rosen's metabolic closure [1] notably expressed as a fixed point of *f(f) = f* where *f* is an operator, operand, and result, and its mathematical realization [2]. The metabolic network is formed by a collection of reactions: A–M->B where A (B) indicates input (output) materials and M is a catalyst. The entire network of reactions must form an operator, operand, and result simultaneously satisfying *f(f) = f* in order to maintain organismic invariance.

Since biological systems are products of an interplay among three inseparable elements: material, energy, and information (entropy), and it is generally accepted that biological systems are information-intensive, we focus on an information aspect of biological closure.

In computation, it had been argued that a certain amount of energy dissipation may be required. However, *reversible computation* [3, 4] has opened up the possibility that energy dissipation could be avoided if computations are carried out in a completely reversible manner by preserving previous states in each step of the computation, since

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 805–812, 2009.

deletion of the information would involve the generation of positive entropy (i.e. negative entropy leak) [4].

Before the reversible computation argument, there had been several arguments on entropy flow within the system. First, it should be noted that von Neumann's interest when building the theory of self-reproducing automata is "building a reliable system from unreliable components" [5]. Schrödinger noted that life consumes negative entropy [6], and we will also relate system reliability to entropy. Bennet, the inventor of the Brownian Machine, noted that "the digital computer may be thought of as engines for transforming free energy into waste heat and mathematical work" [4]. Therefore, a living organism may be characterized as creating "mathematical work" targeted at maintaining itself.

This note revisits self-reproduction [7] and self-reproduction involving errors in description copying [8], and examines biological closure in a context of system reliability involving entropy leak, complexity decrease, and reversibility.

Section 2 examines system reliability, pointing out the similarity between self-reproduction and mutual repairing. Section 3 revisits the self-reproduction model and relates the quasi-species model to the self-reproduction model. Section 4 discusses biological closure and openness based on these models.

2 A System Reliability View of Self-reproduction

2.1 Conventional Reliable Systems

Man-made systems may be characterized as systems whose reliability has been increased by systematically coupling components. Denoting the reliability of the object X as $R(X)$, a system reliability $R(S)$ can be formulated by component reliabilities $R(\mathrm{C_i})$ as:

$$R(S) = \mathrm{f}(R(\mathrm{C_1}), R(\mathrm{C_2}), ..., R(\mathrm{C_N})) \tag{1}$$

where series systems are characterized as:

$\mathrm{f}(R(\mathrm{C_1}), R(\mathrm{C_2}), ..., R(\mathrm{C_N})) = R(\mathrm{C_1})R(\mathrm{C_2}) ... R(\mathrm{C_N})$,

and parallel systems as:

$\mathrm{f}(R(\mathrm{C_1}), R(\mathrm{C_2}), ..., R(\mathrm{C_N})) = 1 - (1 - R(\mathrm{C_1}))(1 - R(\mathrm{C_2})) ... (1 - R(\mathrm{C_N}))$.

Further, these series and parallel systems can be generalized into k-out-of-N where it is series when $k = 1$ and parallel when $k = N$. Similar to the parallel and series arrangement of batteries, any combination of them and any number of hierarchies are possible.

It should be noted that even in man-made systems, system reliability could be made infinitesimally close to one, if one could take an infinite number of components and use them to build a parallel system or infinite hierarchical system, for example.

2.2 Recursive Reliable Systems

When we consider self-repairing using linear sequential thinking, the first stumbling block is the *self-repair paradox* (or equivalently, the self-diagnosis paradox): the system requires a repairing subsystem, which in turn requires another repairing subsystem that repairs the repairing subsystem, thus falling into an infinite regress. The paradox will be resolved in a system that considers self-repair or mutual repair. In [2], the paradox is resolved by avoiding infinite regress with a systemic property (a property dependent on the network connectivity).

With the system reliability formulation above, self-repairing and mutual repairing may be expressed in recursive forms:

$R(\mathrm{C_i}) = g_i\ (R(\mathrm{C_i})\ ,\ R(\mathrm{S}))$ and $R(\mathrm{C_i}) = g_i\ (R(\mathrm{C_i}),\ R(\mathrm{C_j}))$ $(i{\neq}j)$, respectively for the component reliability in (1).

Reproduction and mutual repairing (Fig. 1) are similar in the sense that the former components use offspring for rewinding the probabilistic worn-out and the latter components use themselves being rewound by other components. In fact, system reliability in both cases is formulated in the recursive form of reliability. In repairing, components develop themselves into the future time dimension (Fig. 1, left), but in reproduction they develop themselves to the sample space of their offspring (Fig. 1, right).

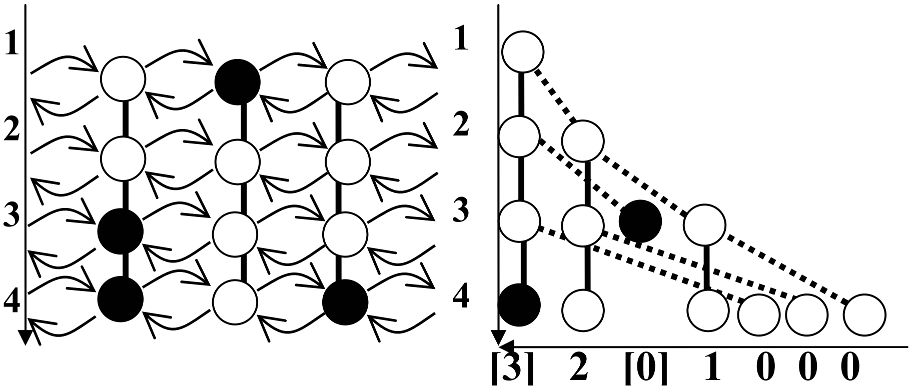


Fig. 1. Diagrams illustrating a mutual-repair with 1-dimensional configuration with periodic boundary condition (left), and self-reproduction (right). The system evolves downward. Time is measured by synchronous mutual-repair trials (left), and by reproduction (right). Arrows indicate mutual repairs. Dotted lines indicate reproduction. Identity of component is indicated by vertical bars. Open nodes indicate normal (alive) components, while black nodes indicate faulty (dead). The horizontal axis of the left figure indicates the age of the current component.

In the mutual repairing system, repairing essentially means overwriting the information of normal components to the degraded information of faulty components. This repairing amounts to “refueling” [4] the faulty components by making their *tape* the ordered one. The action of overwriting information during repair inevitably seems to involve *deletion* of information. Thus, the repairing cannot avoid (negative) entropy leak, or positive *entropy generation* [4].

Self-reproduction also seems to involve entropy leak due to the intrinsic asymmetry of existence and non-existence of materials. Further, when reproduction error is involved, it seems to be even more difficult to avoid entropy leak, for it would make the process irreversible, and we cannot conceive of any way of making it reversible.

The system reliability can be increased further by using the description of the system (as will be explained in the self-reproduction model in Sect. 3.2), but more surprisingly, with room for evolutionability (as will be explained in the quasi-species model in Sect. 3.3).

3 Basic Model

3.1 A Microscopic Model [7]

von Neumann's model consists of $X + \varphi(X)$ where $\varphi(X)$ is a description (genotype) of X (phenotype) [7]. X consists of subsystems *A, B, C*, and *D*. Their functions are defined to map from $\varphi(X)$ to X for A (i.e., a general constructor); and from $\varphi(X)$ to $2\varphi(X)$ (i.e., a copier for description). C controls the copying process. Thus, *A, B*, and *C* are subsystems related to self-reproduction. *D* is the subsystem not involved in the self-reproduction.

As shown in Fig. 2, starting from $X + \varphi(X)$, C makes the copier B copy the description $\varphi(X)$ twice, thereby resulting in three descriptions $3\varphi(X)$. The controller C then activates the general constructor A, and makes A construct X from one description out of three, using the materials in the environment where A is placed. Altogether, the original $X + \varphi(X)$ is doubled. Thus, $X + \varphi(X)$ operates on itself resulting in another $X + \varphi(X)$, which amounts to Rosen's $f(f) = f$.

3.2 Reliability and Complexity

Let *X->Y* indicate that *X* produces *Y*, then the self-reproduction may be expressed by *X->X*. Let *C(X)* further denote a complexity of an object *X*. In the above self-reproduction model, von Neumann restricted himself to the complexity decreasing principle when some automaton makes another automaton. This insight may be formalized in the following property C for complexity and the definition C of the description in terms of the complexity.

***Property C*:** For any production: *X->Y*, $\mathrm{C}(X) \geq \mathrm{C}(Y)$,

***Definition C*:** $\mathrm{C}(\varphi(X)) \leq \mathrm{C}(X)$.

It should also be noted that von Neumann, in his ingenious devising of the self-reproduction *Y->Y* where $Y = X + \varphi(X)$, managed to keep complexity of output not exceeding that of input. It may be conjectured that if the complexity increase or decrease were strict and there could be no neutral way, then reversible reproduction would be difficult and his self-reproduction would inevitably involve entropy leak.

Indeed, in every steps of the self-reproduction this Property C with the Definition C is not violated. In constructing X by A with$\varphi(X)$, $\mathrm{C}(A+\varphi(X)) \geq \mathrm{C}(X)$ holds because $\mathrm{C}(\varphi(X)) = \mathrm{C}(X)$, and in copying the description $\mathrm{C}(B+\varphi(X)\ ) \geq \mathrm{C}(2\varphi(X))$ holds because $\mathrm{C}(\varphi(X)\ ) = \mathrm{C}(2\varphi(X))$.

We can put forth an intuition behind reliability in the following property R and the definition R of a description:

***Property R*:** For a production process X->Y, R(X) ≥ R(Y),

***Definition R*:** R(φ(X)) ≥ R (X)

The property R expresses the simple commonsense that reliability of the product is less than that of the producing system. The definition R is a *reliability version* of the definition of description; the description is an object with material aspect minimized and information aspect kept intact. Thus, description is less subject to degradation when materials undergo degradation. (Another merit of description is that it can utilize a code that would compensate for errors with redundancy introduced by degeneracy in the mapping of the code.) Properties R and C do not contradict each other. On the other hand, the definition C could contradict definition R unless only the equality holds.

To resolve the possible contradiction above or to investigate if there is a contradiction, we need a closer look at the complexity and the reliability. For the complexity, von Neumann extended the above property C and definition C to the followings. However, this again needs further studies such as a quantification of the complexity.

***Property C'*:** For any reproduction: X->Y, C(X) ≤ C(Y) implies that C(X) ≥ ξ for a complexity level ξ.

***Definition C'*:** There is a complexity level π_1 such that C(X) ≤ π_1 implies that C(φ(X)) ≤ C(X), and another level π_2 (≥ π_1) such that C(X) ≥ π_2 implies that C(φ(X)) ≥ C(X).

3.3 A Quasi-Macroscopic Model [8, 9, 10]

Involving the description in reproduction, system reliability can be made higher than that without description. However, using description could involve a description error. Let φ(X)' denote the description including the error. Then the phenotype will be X' which might have a distinct character from X. If we assume the reproduction rate is altered, then we need to focus on the dynamics resulting from the interaction between the population of the original system X + φ(X) and the mutated system X' + φ(X)'.

Let x_1 and x_2 respectively denote the population of the original reproduction system X + φ(X) and the mutant system X' + φ(X)'. Let a_1 and a_2 respectively denote the reproduction rate of the original and the mutant. Further, let R_{11} and R_{22} respectively denote the reliability in reproduction of the original and the mutant, and assuming the failure in reproduction turns out to be another population, that is $R_{12} = 1 - R_{11}$, then $a_2R_{21} = a_2(1 - R_{22}) \simeq 0$ leads to the simplified model of quasi-species [10]:

$x_1 = a_1\, R_{11}\, x_1$, .

$x_2 = a_1(1 - R_{11})\, x_1 + a_2\, R_{22}\, x_2$, where x_i is a time derivative of the variable x_i.

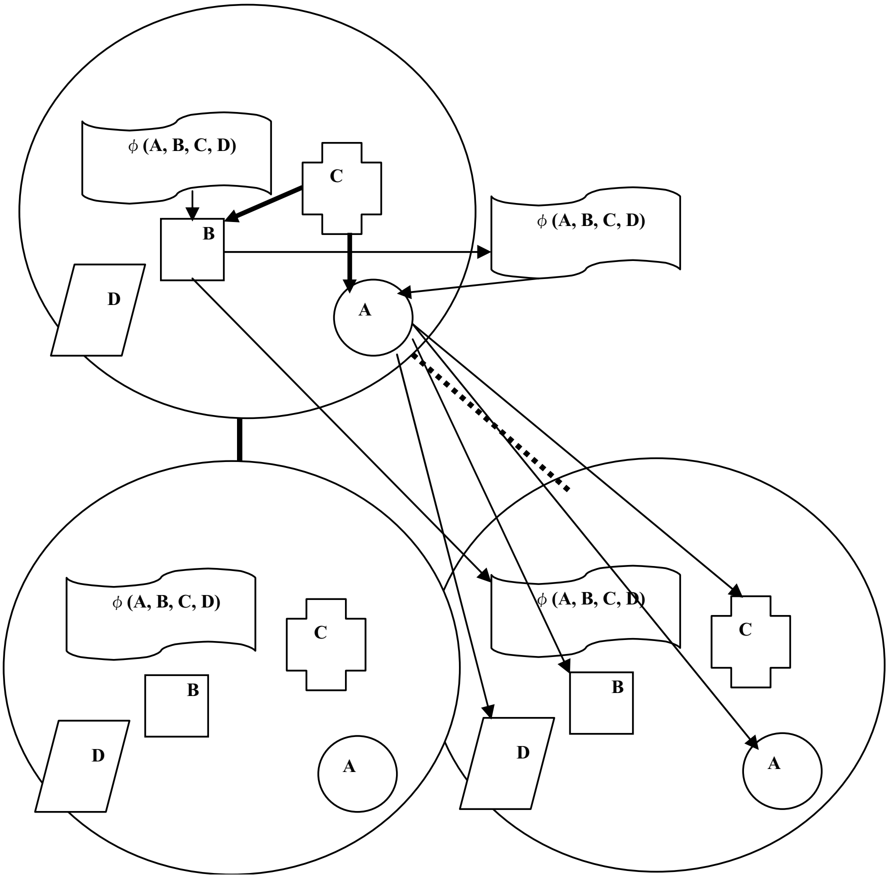


Fig. 2. Diagram magnified on one reproduction in the phylogenic tree of Fig. 1 (left). Arrows indicate input materials and output materials. Thick arrows indicate control signals.

On the one hand, if the reproduction of the original reproduction system is carried out without any error, then there is no mutant and $x_2 = 0$. On the other hand, unless the reliability in reproducing the original one must exceed a certain threshold, that is, (1 >) $R_{11} > R_{22}\ a_2/a_1$, the population of the original one would become extinct. Put another way, if the mutant had the phenotype character of high reproduction rate and reliability such that $R_{22}\ a_2 > R_{11}\ a_1$, it would make the original one extinct. If the reproduction system were to correspond to a cell, then such mutant would correspond to cancer (Fig. 3).

The reliability in self-reproduction is a *double-edged sword*, for if self-reproduction is too reliable (without any error) then the reproduced offspring is a perfect clone of itself that does not involve cancer due to reproduction error, but it also loses the possibility of evolution and adaptation, since there will be no channel to the environment.

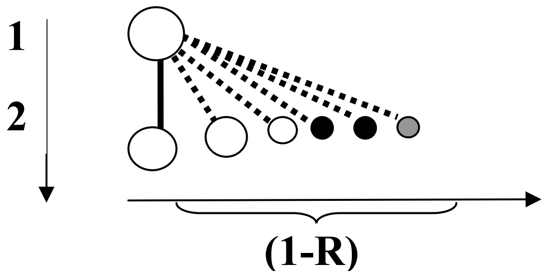


Fig. 3. Quasi-species reproduction involves reproduction errors. Time is measured by generation downward. The vertical axis represents phenotype distance. The size of circles indicates the number of offspring of the type. The gray node indicates a cancer.

4 Summary and Discussion

Inspired by Rosen's (*M*, *R*) system and the mathematical beauty of $f(f) = f$ and its recent realization [2], we revisit biological closure with a bias to information (among three aspects of objects: material, energy, and information). Noting that self-reproduction embodies a mechanism of improving system reliability, and that von Neumann's self-reproduction model realizes the triply entangled fixed point $f(f) = f$, we focused on his model.

The intuition of asymmetry between existence and non-existence on which *reliability* is based implies that any functional device implemented with materials is subject to degradation. This holds true even for descriptions, although the material aspect is minimized. Thus, (negative) entropy leak (i.e. positive entropy generation) seems inevitable from a reliability viewpoint.

von Neumann noted the complexity decrease principle when one automaton produces another. If the complexity decrease were strict, then entropy leak would be supported from the viewpoint of reversibility.

High system reliability attained by self-reproduction does not come for free. For the maintenance of system reliability, the fraction of old (hence degraded) components relative to newly reproduced components must be kept below some ratio. Thus, the system must ensure not only reliable reproduction but also reliable (organized) killing of old components. Although the phenomenon known as *apoptosis* exists in biology, we did not consider controlled death here.

Even higher system reliability may be attained by involving a *description* in self-reproduction. Other than attaining higher reliability than the object being described, the description plays a crucial role in von Neumann's model in managing to keep complexity below a certain level. Otherwise (without the description), an object to object copy must be required, and the complexity of the mechanism of the copy depends on the complexity of the objects being copied. Again, the higher system reliability attained by description-based self-reproduction does not come without a tradeoff. A new type of threat accompanying description-based self-reproduction is due to the description (genotype) error that causes the character of the object being described (phenotype character) of reproduction rate to increase. This threat may correspond to cancer in biology.

Another important aspect accompanying description-based self-reproduction is the diversity in the character of reproduced objects. This diversity opens up the possibility of evolution and adaptation when self-reproducing systems are placed in a community and in a place where interaction with the environment is possible.

Acknowledgments. This note was motivated by Juan-Carlos Letelier's paper [2].

References

1. Rosen, R.: Some realizations of (M;R) systems and their interpretation. Bull. Math. Biophys. 33, 303–319 (1971)
2. Letelier, J.C., et al.: Organizational invariance and metabolic closure: Analysis in terms of (M;R) systems. Journal of Theoretical Biology 238, 949–961 (2006)
3. Fredkin, E., Toffoli, T., Conservative Logic, T.: Conservative Logic. Int. J. Theoret. Phys. 21, 219 (1982)
4. Bennet, C.H.: The Thermodynamics of Computation- a Review. Int. J. of Theoretical Physics 21(12) (1982)
5. von Neumann, J.J.: Probabilistic logics and the synthesis of reliable organisms from unreliable components. In: Shannon, C.E., McCarthy, J. (eds.) Automata Studies, vol. 166, pp. 43–98. Princeton University Press, Princeton (1956)
6. Schrödinger, E.: What Is Life? Cambridge University Press, Cambridge (1944)
7. von Neumann, J.J.: Theory of Self-Reproducing Automata. In: Burks, A.W. (ed.). University of Illinois Press, Urbana (1966)
8. Eigen, M.: Selforganization of Matter and the Evolution of Biological Macromolecules. Naturwissenschaften 58, 465–523 (1971)
9. Eigen, M., Schuster, P.: The Hypercycle. In: A Principle of Natural Self-Organization. Springer, Berlin (1979)
10. Nowak, M.A.: What is a Quasi-species? Trends in Ecology and Evolution 7, 118–121 (1992)

Faith in the Algorithm, Part 2: Computational Eudaemonics

Marko A. Rodriguez[1] and Jennifer H. Watkins[2]

[1] Center for Nonlinear Studies
Los Alamos National Laboratory, Los Alamos NM 87545, USA
marko@lanl.gov
[2] International and Applied Technology
Los Alamos National Laboratory, Los Alamos NM 87545, USA
jhw@lanl.gov

Abstract. Eudaemonics is the study of the nature, causes, and conditions of human well-being. According to the ethical theory of eudaemonia, reaping satisfaction and fulfillment from life is not only a desirable end, but a moral responsibility. However, in modern society, many individuals struggle to meet this responsibility. Computational mechanisms could better enable individuals to achieve eudaemonia by yielding practical real-world systems that embody algorithms that promote human flourishing. This article presents eudaemonic systems as the evolutionary goal of the present day recommender system.

> *[Those who condemn individualism] slur over the chief problems—that of remaking society to serve the growth of a new type of individual.*
>
> John Dewey, “Individualism Old and New”

1 Introduction

Eudaemonia is the theory that the highest ethical goal is personal happiness and well-being [1]. This theory holds that an ethical life is one filled with the meaning and satisfaction that arises from living according to one’s values—where everything one does is of great importance to their character. Eudaemonia parallels the notion of Abraham Maslow’s self-actualization [2] and Mihály Csíkszentmihályi’s flow state [3] except that, as an ethical theory, it argues that it is a personal responsibility to strive for this state. As a social theory, eudaemonia holds that the purpose of society is to promote this state in all of its people. The ethical foundation of personal flourishing is grounded in the contention that the purpose of life is to reap satisfaction and fulfillment from an engagement in the world and that such a state is objectively good for society. Thus, learning how to flourish is a form of moral development.

Moral development, when used in this sense, extends beyond civility, honesty, and other facets of rectitude. It refers to a personal onus to achieve well-being. One proponent of the ethical theory of eudaemonia, David L. Norton, states

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 813–820, 2009.

that "[...] the broader eudaimonistic thesis is that all virtues subsist *in potentia* in every person; thus to be a human being is to be capable of manifesting virtues, and the problem of moral development is the problem of discovering the conditions of their manifestation" [4]. Typically, the discovery of the conditions that will manifest virtues in the individual is guided by the recommendations of family, friends, and community—those who know the individual well and the options available to them. Despite this guidance, the achievement of eudaemonia remains elusive for most. Maslow notes that a very small group of people achieve self-actualization and Csíkszentmihályi has shown that very few are able to control their consciousness well enough to reliably reach the state of flow. Given the individual moral imperative to achieve eudaemonia and the resulting societal benefits, resources should be dedicated to guaranteeing this realization for as many people as possible.

Eudaemonics is the study of the nature, causes, and conditions of eudaemonia [5]. For Owen Flanagan, the domains of moral and political philosophy, neuroethics, neuroeconomics, and positive psychology are the sources from which a developed understanding of human well-being will spring. In this article, it is posited that *computational eudaemonics* will make advances to bring eudaemonia to more than a select few in society. Computer and information science can greatly contribute to the eudaemonic endeavor by yielding practical real-world systems that embody algorithms that promote human flourishing. Systems that promote eudaemonia are called eudaemonic systems. Such systems would foster eudaemonia by providing the right conditions for the manifestation of virtues. This article presents a vision of eudaemonic systems as the evolutionary goal of the present day recommender system.

2 From Recommender to Eudaemonic Systems

The purpose of a eudaemonic system is to produce societies in which the individuals experience satisfaction through a deep engagement in the world. This engagement can be fostered by uniting individuals with those resources that meet their needs. Resources can take many forms, a few of which are itemized below.

- activities: vocations, hobbies, gatherings, projects.
- education: universities, lectures, areas of study.
- entertainment: books, movies, music, shows.
- people: friends, work associates, life partners.
- places: to live, to vacation, to dine.

There are many ways a eudaemonic system could contribute to individual well-being. Perhaps the most ambitious eudaemonic system is one that supplies the satisfaction of the need for a resource before the need is even felt. For Thomas Hobbes, eudaemonia is encumbered by conation—goals, plans, and desires [6]. Practically speaking, humans seek books and movies to stimulate their cognitive faculties, friends and partners to fulfill their social affinities, art to entice their affective natures, and sports to satiate their physical needs. While every individual longs for varying degrees of these requirements, in general, a flourishing life is one where all these requirements are met through the active process of enacting them [7]. Thus, a Hobbesian eudaemonic system would be one that satisfied requirements before they were felt (pre-conation), so that the experience of need could not disrupt a life of contentment. Through computational mechanisms, it may be possible to produce pre-conate eudaemonic systems. A pre-conate system is one that makes use of indicators of coming discontent and provides avenues to rectify the situation prior to its actualization.

Recommender systems [8], when viewed within the context of the eudaemonic thesis, could evolve to become such systems. A recommender system is an information filtering tool that matches individuals to resources of potential interest. Such systems are commonly employed by businesses in an attempt to sell more products. However, this conceptualization of the recommender system trivializes their potential role.

The satisfaction one reaps from the world can be represented in terms of one's interactions with resources. These interactions need not be extraordinary, but are the stuff of everyday life. Norton articulates the importance of everyday activities when he states that "if the development of character is the moral objective, it is obvious that [...] the choices of vocation and avocations to pursue, of friends to cultivate, of books to read are moral for they clearly influence such development" [4]. For the techno-social society, this development of character is driven every day, to some extent, by the use of recommender systems. Thus, to the extent that recommender systems influence choices, they already influence moral development. By purposely designing these systems to orient individuals toward life optima, recommender systems can evolve to become eudaemonic systems.

The current generation of recommender systems are limited to a particular representational slice of the world (such as movies). This is represented in Figure 1a, where there exists a tight coupling between the data and the application which operates on that data. A eudaemonic system must account not for a single aspect of an individual's life, but for the multitude of domains in which that individual exists. The emerging Web of Data provides a distributed data structure that cleanly separates the data providers from the application developers. This is represented in Figure 1b. The remainder of this section will discuss recommender systems and their transition to eudaemonic systems through the exploitation of the Web of Data.

2.1 Recommender Systems

Most recommender systems model individual users, resources, and their relationships to one another [8]. For example, in an online store, users may have an `ex:hasPurchased` relationship to some of the store's products. If the purchasing behavior of user x and user y has a strong, positive correlation, then any products purchased by only one can be recommended to the other. Purchasing behavior is not the only way in which resources are deemed similar. It is

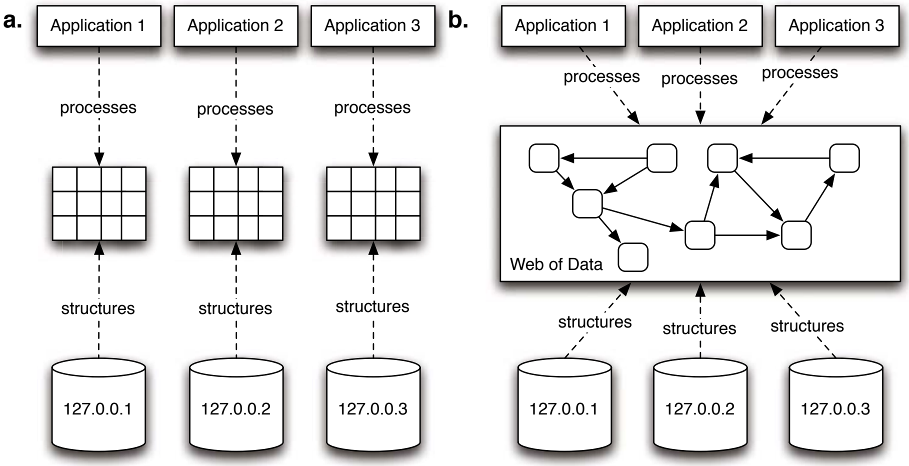


Fig. 1. a.) The current paradigm in which the application and the data upon which it operates are tightly coupled both technically and proprietarily—§2.1. b.) The emerging Web of Data provides a collectively generated, publicly accessible world model that can be leveraged by independent application developers—§2.2.

possible to relate resources by shared metadata properties [9]. For example, an online movie rental service can represent movie a as having an `ex:directedBy` relationship to director b and director b can maintain an `ex:directed` relationship to movie c. The similarity that exists between movies a and c is determined, not by user behavior, but by similarity of metadata—the same person directed both. By building a graph of typed relationships between resources, it is possible to identify different forms of relatedness and utilize these forms to aid an individual in their decision making process regarding the use of such resources.

The power of recommender systems is currently limited because they rely on a single silo of data that must be generated before they can provide useful recommendations (see Figure 1a). Due to the data acquisition hurdle, application designers must focus on a particular niche in which to provide recommendations. For example, services either provide recommendations for books,[1] or for music,[2], or for partners,[3] etc. With such a limited worldview, these services do not respect the multi-faceted nature of human beings. If a system only has access to data on movies, then it can never recommend the perfect beach novel. Eudaemonia requires a complete representation of the domains in which one conducts life in order to recommend the right resource at the right time. Therefore, eudaemonic systems require an integrated representation of the world's resources and the individual's place within them.

[1] For example: Amazon.com, Feedbooks.com

[2] For example: Pandora.com, Last.fm

[3] For example: Match.com, Chemistry.com, eHarmony.com

2.2 Eudaemonic Systems

The recommender system data structure described previously can be conveniently represented as a multi-relational network. The most prevalent multi-relational data model is the Resource Description Framework (RDF). The Linked Data community is dedicated to the development of the emerging RDF-based Web of Data. On the Web of Data, all data is represented in the URI address space and interlinked to form a single, global data structure that can be used by both man and machine for various application scenarios (see Figure 1b) [10].[4] The Web of Data provides two significant benefits over the data silos used by recommender systems. First, application developers need not focus on data acquisition and instead can focus directly on algorithm development. This feature ultimately reduces the labor involved in web service deployment. Second, the application developer can create algorithms that make use of a rich world model that incorporates the various ways in which resources relate to each other. Thus, these algorithms have a larger knowledge-base with which to understand the world and the individual's place within it.

Figure 2 presents a visualization of the linking structure of the 89 data sets currently in the Linked Data cloud.[5] Each vertex represents a unique data set that exists on an Internet server. The directed relationships denote that the source data set references resources in the sink data set. The current Linked Data cloud maintains approximately 4.5 billion relationships on data from various domains of interest. Table 1 indicates the domain of interest for each data set. By publicly exposing data sets such as Amazon.com's RDF book mashup, MusicBrainz.org's metadata archive, the Internet Movie Database's (IMDB) collection of movie facts, Revyu.com's user ratings, and the publishing and conference behavior of scholars, the Web of Data hosts a rich model of the world that is not built by a single provider, but by many providers collaboratively integrating their data. Such a massive public data structure can be exploited by a community of developers focused on ensuring that the right resource reaches the right person at the right time.

The Web of Data already includes data sets that are pertinent to modeling individuals and resources; however, the success of a eudaemonic system depends on the availability of data regarding the individual and their past, current, and predicted responses to resources. At the societal level, research has demonstrated that resources relevant to flourishing are those that support life expectancy, nutrition, purchasing power, freedom, equality, education, literacy, access to information, and mental health [12].[6] At the individual level, gathering and

[4] The public exposure of data has stimulated interest in the development of the legal structures for the use of such data. Much like the Open Source movement, the Linked Data community is actively involved in the Open Data movement [11].

[5] The Linked Data cloud is a subset of the larger Web of Data that includes those data sets that are directly or indirectly connected to DBpedia and are maintained by the Linked Data community.

[6] The World Database of Happiness provides data concerning the study of well-being worldwide and is available at `http://worlddatabaseofhappiness.eur.nl`

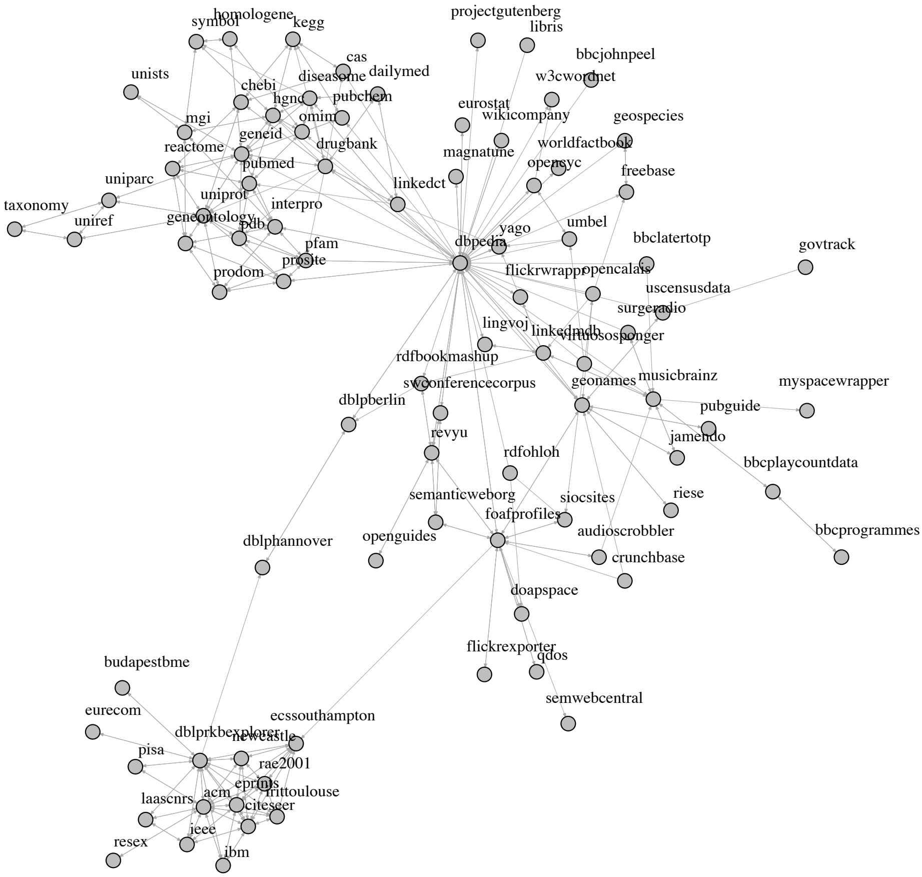


Fig. 2. A representation of the 89 RDF data sets currently in the Linked Data cloud

maintaining data regarding fluctuations in an individual's well-being in relation to resources would support the automatic determination of optimal future states for that individual. The algorithms that define this automatic determination will be constrained by the same validation requirements seen in today's recommender systems—e.g. precision and recall, and more tellingly, use. Eudaemonic algorithms must be able to adapt to user requirements which continually change with habituation and personal development. Those algorithms that do not adapt to the changing user will simply not be adopted.

While the Linked Data community is providing a distributed data structure, they are not providing a distributed process infrastructure [13]. Currently, the Linked Data practice is to mint `http`-based URIs. These `http`-based URIs are dereferenced in order to retrieve a collection of RDF statements associated with that URI. The problem with this model is that the Web of Data is primarily useful to man, not machine. For a machine to traverse parts of the larger Web of Data, the pull-based mechanism of HTTP greatly reduces the speed of

Table 1. The domains of the 89 data sets currently in the Linked Data cloud

data set	domain	data set	domain	data set	domain
acm	computer	geospecies	biology	pubchem	biology
audioscrobbler	music	govtrack	government	pubguide	books
bbcjohnpeel	music	hgnc	biology	pubmed	medical
bbclatertotp	music	homologene	biology	qdos	social
bbcplaycountdata	music	ibm	computer	rae2001	computer
bbcprogrammes	media	ieee	computer	rdfbookmashup	books
budapestbme	computer	interpro	biology	rdfohloh	social
cas	biology	irittoulouse	computer	reactome	biology
chebi	biology	jamendo	music	resex	computer
citeseer	computer	kegg	biology	revyu	reference
crunchbase	business	laascnrs	computer	riese	government
dailymed	medical	libris	books	semanticweborg	computer
dblpberlin	computer	lingvoj	reference	semwebcentral	social
dblphannover	computer	linkedct	medical	siocsites	social
dblprkbexplorer	computer	linkedmdb	movie	surgeradio	music
dbpedia	general	magnatune	music	swconferencecorpus	computer
diseasome	medical	mgi	biology	symbol	medical
doapspace	social	musicbrainz	music	taxonomy	reference
drugbank	medical	myspacewrapper	social	umbel	general
ecssouthampton	computer	newcastle	computer	uniparc	biology
eprints	computer	omim	biology	uniprot	biology
eurecom	computer	opencalais	reference	uniref	biology
eurostat	government	opencyc	general	unists	biology
flickrexporter	images	openguides	reference	uscensusdata	government
flickrwrappr	images	pdb	biology	virtuososponger	reference
foafprofiles	social	pfam	biology	w3cwordnet	reference
freebase	general	pisa	computer	wikicompany	business
geneid	biology	prodom	biology	worldfactbook	government
geneontology	biology	projectgutenberg	books	yago	general
geonames	geographic	prosite	biology		

processing. It would be unfortunate to limit the sophistication of the algorithms that can reasonably process this data due to an infrastructure issue that can be solved using distributed computing.

In addition to issues of technical implementation, eudaemonic systems present a number of social concerns. For example, a system that provides individuals with exactly what they want would erode personal motivation to achieve. However, eudaemonic systems, by definition, cannot encourage such “lotus eating” as self-indulgence does not lead to flourishing. Csíkszentmihályi defines the state of psychological flow as the balance between simplicity (boredom) and complexity (frustration) [3]. Thus, the intent of a eudaemonic system is to match the individual with resources that push their cognitive limits, where goals are challenging, but not impossible. It is, in fact, the sole purpose of a eudaemonic system to ensure that individuals realize continual, life-long personal achievement.

3 Conclusion

The evolution of the recommender system to the eudaemonic system will be driven by the public exposure of massive-scale, interlinked, heterogenous data and algorithms that can effectively and efficiently process such data. The goal of a eudaemonic system is to orient people towards those resources that will produce a life that is devoid of pretense, doubt, and ultimately, fear. That is, a eudaemonic system will aid the individual in situating themselves within that area of the world that makes sense to them. A pre-conate eudaemonic system would direct the individual to choose need-mitigating options before the individual becomes aware of their need. In other words, the individual would choose options that they do not perceive as necessary. Without the perception of need, the individual would take on faith that the algorithm knows what is best for them in a resource complex world. Thus, the perfect life is not an aspiration, but a well-computed path.

Note

Faith in the Algorithm is a series of articles that focuses on the intersection of political philosophy, ethics, and computation.

References

1. Aristotle: Nicomachean Ethics (350 B.C.)
2. Maslow, A.H.: A theory of human motivation. Psychological Review (50), 370–396 (1943)
3. Csíkszentmihályi, M.: Flow: The Psychology of Optimal Experience. Harper and Row, New York (1990)
4. Norton, D.L.: Democracy and Moral Development: A Politics of Virtue. University of California Press (1995)
5. Flanagan, O.: The Really Hard Problem: Meaning in a Material World. MIT Press, Cambridge (2007)
6. Hobbes, T.: Leviathan (1651)
7. Kraut, R.: What is Good and Why: The Ethics of Well-Being. Harvard University Press (2007)
8. Resnick, P., Varian, H.R.: Recommender systems. Communications of the ACM 40(3), 56–58 (1997)
9. Pazzani, M.J., Billsus, D.: Content-Based Recommendation Systems. In: Brusilovsky, P., Kobsa, A., Nejdl, W. (eds.) Adaptive Web 2007. LNCS, vol. 4321, pp. 325–341. Springer, Heidelberg (2007)
10. Bizer, C., Heath, T., Idehen, K., Berners-Lee, T.: Linked data on the web. In: Proceedings of the International World Wide Web Conference. Linked Data Workshop, Beijing, China (April 2008)
11. Miller, P., Styles, R., Heath, T.: Open data commons: A license for open data. In: Workshop on Linked Data on the Web. ACM Press, New York (2008)
12. Heylighen, F., Bernheim, J.: Global progress I: Empirical evidence for increasing quality of life. Journal of Happiness Studies 1(3), 323–349 (2000)
13. Rodriguez, M.: A Reflection on the Structure and Process of the Web of Data. Bulletin of the American Society for Information Science and Technology 35(6), 38–43 (2009)

System Engineering Security

Esmiralda Moradian

Department of Computer and System Sciences,
Stockholm University
Forum 100, 164 40 Kista, Stockholm, Sweden
esmirald@dsv.su.se

Abstract. Organizations' integrate different systems and software applications in order to provide a complete set of services to their customers. However, different types of organisations are facing a common problem today, namely problems with security in their systems. The reason is that focus is on functionality rather than security. Besides that, security, if considered, comes too late in the system and software engineering processes; often during design or implementation phase. Moreover, majority of system engineers do not have knowledge in security. However, security experts are rarely involved in development process. Thus, systems are not developed with security in mind, which usually lead to problems and security breaches. We propose an approach of integration security throughout engineering process. To assure that necessary actions concerning security have been taken during development process, we propose semi-automated preventive controls.

Keywords: System engineering, software development, security, risk management, control, security breach.

1 Introduction

Enterprises, governmental, medical, and defence sectors are dependent on software developed by system engineers. Software weaknesses and defects can result in software application failure, but also be exploited by attackers. [16]. Software is everywhere [16]. Therefore, a poorly designed and developed as well as not properly tested software will induce security breaches that, for example, can cause economical problems, and/or affect humans' health and, in the worse case, even cause death. Thus, systems that handle sensitive, secret and/or valuable information must operate correctly and have a stable status. Organisations' systems handling sensitive data are attractive targets for attackers. Despite of a large number of different security standards and mechanisms and tools existing today, vulnerabilities in systems are still present. During the recent years we could observe criminal activities performed on so called "secure systems" as well as errors in software. One example is bank systems. During few years direct aimed attacks were performed on Nordea bank's systems. Another example is the medical sector. Due to software error in the Therac-25, caused by software bug many people received radiation overdoses that affected their lives [16]. Vulnerabilities are usually the result of defective specifications, design,

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 821–828, 2009.

implementation, and testing that are unknowingly injected into software by system developers. [13] To a large extent flaws and defects can be minimized. However, to provide system reliability, correct operation and precise performance - security aspects should be considered early in system development lifecycle. Systematic and structured actions such as risk management (identification, analysis, evaluation of risks as well as risk treatment, monitoring and review) and threat modelling are required. Authors in [20] presented the survey made by UK Department of Trade and Industry. According to the survey many companies do not have sufficient internal controls of systems on the Internet. Organisations are often satisfied with "secure" software they purchase for these reasons.

Unfortunately, many organisations are not considering software security during the development life cycle. Commonly, security is added sometimes after the system is developed. Bishop in [3] states that systems where security mechanisms were added after the system was built are not trustworthy. We propose an approach where security is interspersed in the system engineering process. Semi-automated control of performed security measures during different phases of software development life cycle (SDLC) can facilitate building necessary and required security. The approach involves not only computer specialists but also managers/decision makers who have responsibility for the information security, and customers who present requirements.

2 Related Work

Mouratidis, et al. [14] argues for the need to develop a methodology that considers security as an integral part of the whole system development process. Authors propose an approach that considers security concerns as an integral part of the entire system development process. The different stages of the approach are described with the help of a health and social care information system.

N. Haridas in [8] point out that application developer never consider or have a disciplined process to address security in any phases of SDLC. The author proposes incorporating security in different SDLC phases. N. Haridas [8] presents security factors followed by an example. Further the author presents a list of prioritized factors that could be considered as security guideline during the development phase

G. McGraw in [13] proposes building security in SDLC. The author presents a detailed approach for putting software security into practice. G. McGraw [13] provides guidance on how to build secure software and shows gaps in the development process and to how to improve the process. However, while there exist different approaches and guidance's on how to integrate security in system development process, decision makers and developers hardly adopt those. We propose approach of integration security throughout engineering process with support of semi-automated and preventive controls in each phase of SDLC process.

3 Security Issues in Development Life Cycle

"The essence of good security engineering is understanding the potential threats to a system, and then applying an appropriate mix of protective measures – both technological and organisational – to control them" [1]. In many organisations managers and developers teams start, in the best case, thinking about security when system is already shipped in production, which affects quality, security and economics. Vulnerabilities embedded in software and system components affect security of the system. However, as a result of vulnerability, security breaches can be costly both in terms of effort to fix problems and damage to organisations. For example security breaches cost more money, take more time and can bankrupt a company and destroy information. (See Figure 1)

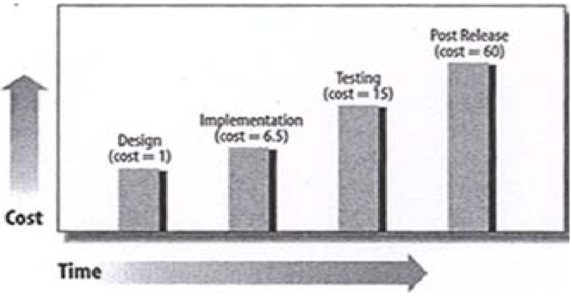


Fig.1. Cost of fixing security flaws during different development phases Source: Secure Coding – Principles&Practices Mark G. Graff, Kenneth R. van Wyk O'Reily

Moreover, adding security afterwards can impact people's everyday's life. Customers and end users are rare aware of software insecurity and will expect systems, applications, products and services to be secure. [9, 18] Development process deals with transformation of requirements into a software product or software-based system that meets the customer's defined needs. Gathering and capturing requirements is one of the critical parts of development process, which affect system development during all other phases. Requirements express behaviour of the system, the system and object states and the transition from one state to another. Requirements also describe system's activities. Usually, requirements are described as functional and non-functional [15]. The purpose of software and system requirements analysis "is to establish the requirements of the software elements of the system" and "transform the defined stakeholder requirements into a set of desired system technical requirements that will guide the design of the system" [11]. However, it is common that security requirements of software and system are not even identified. During the design developers trying to identify "which system requirements should be allocated to which elements of the system." [11] Software design is design that implements and can be verified against requirements. However, security requirements that are not identified and analysed will lead to security flaws in system architecture and design. Testing ensures that the implementation of each system requirement is tested and that the system is ready for delivery [11]. Though again, missing security requirements, security flaws in architecture and design will produce unreliable test results. Security testing should be a part of development life cycle. In this paper we propose integration of security throughout engineering process.

Several standards, procedures, methods, and tools was developed in purpose to help organisations to understand, develop and implement software systems as well as manage and control organisations' processes. However, it seems to creates problems "in management and engineering, especially in integrating products and services" [11]. For example, C.B. Haley in [7] argues that ISO and NIST documents "provide little guidance on how to connect the functionality to the security needs. Instead of describing when and why objects are to be protected, they describe how the objects are to be protected." Existing standards for information and software provide guidance and support for organisations. For example, ISO 12207 provide guidance to software life cycle architecture [11]. ISO 27001[10] provide a "Plan-Do-Check-Act model" to structure the organizations processes. The purpose is establishing, implementing, maintaining, and improving information security management system [10]. ISO 27002 provide guidance for "initiating, implementing, maintaining, and improving information security management within an organization" [10]. ISO/IEC 15026 [22] defines a process for establishment of integrity levels. Risk analysis and system architecture analysis are the part of the process. COSO-ERM (Enterprise Risk Management - Integrated Framework) was created to help businesses and other entities assess and enhance their internal control systems [6]. Common Criteria (CC) is a framework for evaluating security products and systems. [5] However, many organisations are not working according to one or more security standards because of different reasons. One reason can be difficulty to choose the most suitable standard for organisation. As appear from C. Magnusson's [12] report there exist gaps and overlaps in standards [12]. Author studied five standards, among others ISO 27001 and COSO-ERM.

4 Development and Control

To build secure software system requires effort from all parties: managers, architectures, designers, developers, testers [21], as well as customers and consumers. Security experts are seldom involved in the process; consequently security is left aside. Moreover, people involved in software system development process often go beyond their expertise and do work in other domain areas. That causes problem but seems to be the rule rather than exception in software development. [19] Software and system engineers, usually, are experts in one or more area, such as software architecture, programming, testing. The Engineer's expertise comprises the development process and maintenance process. [19] The managers' expertise comprises organisational life cycle process that for example includes initiation and scope definition, planning, review and evaluation, and closure. [19]. Managers, developers, and customers are usually not experts in security and have different understandings and views on it. This can create misunderstanding and consequently lead to wrong decisions that impact the result. It is important to emphasise that a system probably will be attacked if it has valuable assets that attract attackers and entry points. Thus, to be able to build secure enough systems it is necessary to involve security experts early in the development process.

To provide reliable system and/or service software life cycle and security life cycle should be joined and melt together. Our approach is intersperse security in the system engineering process. Software engineering process lacks visibility and continuity. It is difficult to see progress in software construction. [19] To see and track decisions, implemented security measures, and mechanisms, semi-automated control of performed actions, such as, checks, logs and feedbacks are required in every phase of system engineering process. To assure incorporation of security in the life process, we propose the semi-automated control regardless organisation's branch, development process model, techniques and tools used. Semi-automated control with tracking of performed actions during different phases of software engineering process can facilitate in building necessary and required security (see Figure 2). Security activities are summarised according to classic V-model. Activities in the left side are concentrated on security requirements and secure design, while activities on the right side focus on verification and validation throughout the entire software life cycle. We propose to start from managers, including security managers that make decision to start the project/process of software development. Those need to classify level of security depending on different inputs. Examples of inputs are: business goals and conditions, policies, standards, environment where system should operate, requirements, and knowledge and expertise, and limitations. (See Figure 2)

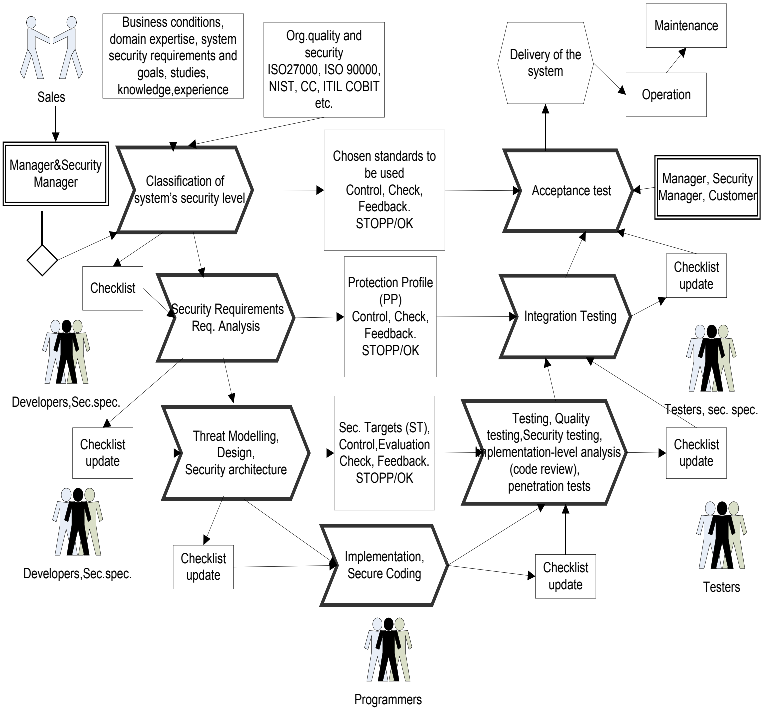

Sales
Business conditions, domain expertise, system security requirements and goals, studies, knowledge,experience
Org.quality and security ISO27000, ISO 90000, NIST, CC, ITIL COBIT etc.
Delivery of the system
Maintenance
Operation
Manager&Security Manager
Classification of system's security level
Chosen standards to be used Control, Check, Feedback. STOPP/OK
Acceptance test
Manager, Security Manager, Customer
Checklist
Security Requirements Req. Analysis
Protection Profile (PP) Control, Check, Feedback. STOPP/OK
Integration Testing
Checklist update
Developers,Sec.spec.
Testers, sec. spec.
Checklist update
Threat Modelling, Design, Security architecture
Sec. Targets (ST), Control,Evaluation Check, Feedback. STOPP/OK
Testing, Quality testing,Security testing, Implementation-level analysis (code review), penetration tests
Checklist update
Developers,Sec.spec.
Checklist update
Implementation, Secure Coding
Checklist update
Testers
Programmers

Fig. 2. Controlled Development life cycle

Depending on application, security level will vary. The security level of bank system that handles electronic payment through online and Internet banking will be much higher than security level of system that handles parking tickets. However, it is important to point out that every software system should have basic level of security, i.e. security baseline. Risk assessment should also be done. Based on classification of the security level, system generates an output – automated security guideline and control.

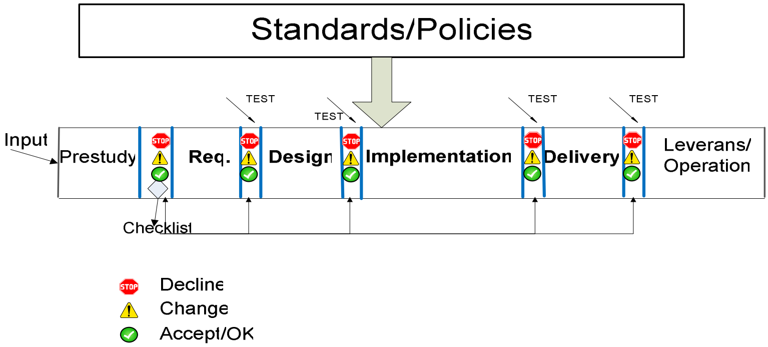


Fig. 3. Control checkpoints in system development process

It can be constructed as checklists with multiple checkpoints. We use software agents for this purpose. These checkpoints should be considered and implemented during development process. The checklist can contain points concerning information security, secure design, network security, and security testing. Some examples of checkpoints are: random password generation, access control, error messages not reveal too much information to an attacker, threat analysis, penetration tests. If some points were not implemented or changed during the process an explanation should be present in updated on each phase checklist. Checks are performed on every phase of development process and all actions are monitored and verified. (See Figure 2 and Figure 3) System will record all changes in checklist, track users and record their actions during the process. Thus, checklist can be considered on the one hand as support for managers, developers and testers in decision making and on the other hand as control of actions taken.

Moreover, ability to provide trusted system or service requires dependability. Allen et al. state that dependable software is software that will "never deviate from correct operation under anticipated conditions" [4]. Dependability requirement of the system can be different, depending on application. [2] Dependability encompasses following attributes: reliability, safety, security (with respect to confidentiality, integrity, availability) and maintainability [2] and sets requirements on the entire system [21]. Depending on the security requirements, protection profile (PP), specific actions will be required to be taken by system developers. System will track decisions made. Decision(s) should be evaluated by evaluator and feedback given about "whether a PP is complete, consistent, and technically sound and hence suitable for use in developing an ST" (Security Targets). [5] Risk management process is an essential part in implementing adequate security level [17]. It includes communication, environment (intern and extern) and risk assessment. Risk assessment consists of risk identification, analysis and evaluation. The purpose of risk assessment is to produce knowledge about relevant security characteristics of systems. Identifying valuable assets and performing risk assessment and threat modelling are the necessary actions to be completed before implementation phase. Pfleeger [15], Swiderski and Snyder [18] points out that organization must understand the vulnerabilities to which it may be exposed. Vulnerabilities can be determined by performing risk analysis. Moreover, risk analysis includes threat identification, analysis and modelling. These can describe when and why objects need protection. The authors in [3] point out two types of risk analysis methods, namely quantitative and qualitative methods. However, to perform proper risk analysis integration of both is necessary.

Design should include designing and modelling controls that detect and/or prevent risks identified in previous phase. This stage also includes modelling rules that will be enforced in implementation phase. During design phase all stated requirements should be satisfied. Otherwise the process is stopped and cannot continue. Desired security behaviour should be in balance with functionality requirements. Control check list must be updated. All changes should be documented. Implementation stage is about to transform design into one or more programming language and assure that system is working as expected. Testing is done iteratively in order to verify specified requirements and validate that all requirements for a specific intended use of the software work product are fulfilled. Security testing is necessary in order to verify that the system design and code can stand against attack [9]. If or when security flaw is found in the design or in code it should be fixed, recorded and analyzed. Security testing is about determining if features work in different way than it was anticipated by developer(s) [9].

5 Conclusion and Further Work

Security in system engineering plays an increasingly important role in nowadays business. Security has impact core business processes in every organization. We have emphasized some important problems in development process in order to enlighten the managers and the developers about the gaps in communication as well as common lack in visibility and continuity that have impact on security of system engineering. We proposed an approach of integration of security in each phase of system engineering process and implementation of semi-automated control in order to log changes in checklist, track users and record their actions during the process. However, the outcome will not exclude earlier made bad decisions that affect the continuous decisions and actions in system developing process. The proposed approach can enable qualitative, secure and effective way of system development.

The next step is to examine every step in detail, starting from requirement analysis.

References

1. Andersson, R.: Security Engineering A guide to building Dependable Distributed Systems, 2nd edn.
2. Avizienis, A., Laprie, J.-C., Randell, B.: Fundamental Concepts of Dependability. UCLA CSD Report no. 010028 LAAS Report no. 01-145 Newcastle University Report no.CS-TR-739
3. Bishop, M.: Introduction to Computer Security. Pearson Education, Inc., London (2005)
4. Allen, J.H., Barnum, S., Ellisson, R.J., McGraw, G., Mead, N.: Software Security Engineering A Guide for Project Managers. Addison-Wesley, Reading (2008)
5. CC, 2006. Common Criteria for Information Technology Security Evaluation, Part 1: Introduction and General Information. Version 3.1 Revision 1 (September 2006)

6. COSO-ERM Enterprise Risk Management —Integrated Framework. Executive Summary (September 2004)
7. Haley, C.B., Moffett, J.D., Laney, R., Nuseibeh, B.: A Framework to security requirements engineering. In: SESS 2006, Shanghai, China, May 20–21, 2006. ACM, New York (2006)
8. Haridas, N.: Software Engineering – Security as a Process in the SDLC, April 2, 2007. SANS Institute InfoSec Reading Room (2007)
9. Howard, M., LeBlanc, D.: Writing Secure Code, 2nd edn. Microsoft Press (2003) ISBN 0-7356-1722-8
10. http://www.27000.org/
11. http://www.12207.com/
12. Magnusson, C.: Corporate Governance, Internal Control and Compliance (September 2007), http://www.svensktnaringsliv.se/material/rapporter/article35898.ece
13. McGraw, G.: Software Security Building Security in. Addison-Wesley, Pearson (2006)
14. Mouratidis, H., Giorgini, P., Manson, G.: When security meets software. engineering: A case of modelling secure information systems (2005) ISSN: 0306-4379
15. Pfleeger, S.L.: Software Engineering Theory and Practice, 2nd edn. Prentice-Hall, Inc., Englewood Cliffs (2001)
16. Rice, D.: Geekonomics The Real Cost of Insecure Software. Pearson Ed. Inc., London (2008)
17. Sherwood, J., Clark, A., Lynas, D.: Enterprise Security Architecture A Business-Driven Approach. CMP Books (2005) ISBN 1-57820318-X
18. Swiderski, F., Snyder, W.: Threat Modelling. Microsoft Press (2004) ISBN 0-7356-1991-3
19. Van Vliet, H.: Software Engineering Principles and Practice, 2nd edn. John Wiley and sons, Chichester (2004)
20. Boer, T., Booijink, T., Liezenberg, C., Nienhuis(Innopay), J.J., Bryant, C., Pruneau(EBA), A.: E-invoicing 2008 European market description and analysis. V. 1.0 February 2008 Copyright © 2008 Euro Banking Association (EBA) and Innopay
21. Lindström, C., Näsström, S.: Handbook for Software in Safety-Critical Applications. Swedish Armed Forces (2005)
22. http://www.iso.org

A Semantically-Based Task Model and Selection Mechanism in Ubiquitous Computing Environments

Angel Jimenez-Molina, Jun-Sung Kim, Hyung-Min Koo, Byung-Seok Kang, and In-Young Ko

Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea
{anjimenez,junkim,hmkoo,byungseok,iko}@kaist.ac.kr

Abstract. User centricity required by ubiquitous computing is about making services and information be prepared and delivered in the perspective of users rather than system elements. Task-oriented computing supports user centricity by representing users' goals in tasks. It bridges the gap between tasks and available services. This paper proposes a semantically-based generic model for describing tasks in ubiquitous computing environments. This model is used by a task selection algorithm that considers the context information of a user and the surrounding environment. Additionally, this paper proposes a pattern-based task reconfiguration algorithm. The algorithms are illustrated by a demo application conducted in our test bed, and by other examples of tasks selected under diverse situations. Evaluation results show a reasonable time overhead for the task selection algorithm.

Keywords: Ubiquitous Computing, Task-oriented Computing, Semantic Web.

1 Introduction

Ubiquitous Computing (Ubicomp) is a paradigm shift that enables users to access networked services everywhere and at anytime [1]. It leads to an invisible technology that disappears from users' consciousness, while supporting their daily life routines. This challenge requires user centricity, which is about making services and information be prepared and delivered in the perspective of users' goals rather than system elements.

The approach of task-oriented computing supports user centricity. This approach realizes the perspective of users by representing their goals in tasks, which are then mapped to available services in the Ubicomp environment based on context information [2]. The context information characterizes the state of the users and their surrounding Ubicomp environment. Therefore, task-oriented computing is about bridging the gap between tasks and services, taking into account their different perspectives, granularities and abstraction levels [3], [4]. The research in task-oriented computing has tended to focus on selecting and integrating services that meet users' task requirements [5]. Moreover, the research has emphasized statically binding a set of services with a task by using predefined rules, thus reacting in the way the rule was designed [6].

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 829–837, 2009.

This paper defines a mediating process composed of three composition layers. In the *task layer,* tasks are defined by a set of coordinated actions, named as an application template. In the *composition-pattern layer,* each action is supported by a configuration of abstract services, represented in a service composition pattern (SCP). A SCP maps an action to the *service layer*, composed of service instances that are tightly bound to the local devices and Web-based computational resources.

Ubicomp environments are highly dynamic due to frequent changes in users' goals and context information, which require a runtime reconfiguration of tasks. Therefore, the aim of this paper is in two folds – first, to provide a flexible task selection mechanism based on the context information; and second, to provide a reconfiguration mechanism based on changing context information. The requirements considered in the design of these mechanisms are as follows: (1) the essential characteristics of users and a Ubicomp environment need to be reflected in the description of tasks; (2) the contextual information gathered from the environment needs to be interpreted in terms of the domain characteristics, and (3) changes on context information require reconfiguration of tasks during runtime by reconsidering the set of SCPs that support their actions.

In order to meet these requirements, we have developed a semantically-based task description model. Essential semantic elements of this model are taken from a real dataset, known as a Time-use Study [7]. Partridge et. al. [8] identified an essential set of contextual variables, and showed that the use of the Time-use Study data is effective to recognize significant user activities in 80% accuracy. We have also developed a semantic and priority-based task selection algorithm, as well as a pattern-based task reconfiguration algorithm based on this model. The algorithms are illustrated by showing an example in a Ubicomp home environment. Additionally, experimental results over a real Time-use Study dataset show a suitable time overhead of the task selection mechanism.

The plan of this paper is as follows. Section 2 provides a semantic representation model of tasks and actions. The mechanisms are explained in detail in section 3. Section 4 describes the implementation, experiments with a demo scenario and other examples, and a performance evaluation. Section 5 provides related work and section 6 concludes the paper.

2 Semantic Representation Model for Tasks and Actions

Task Properties. Recently the Ubicomp community has realized the potential benefits of supporting task recognition mechanisms by understanding users' behavior from real datasets. During decades humans' daily tasks have been recorded in publically available datasets, named Time-use Studies [7]. These tasks are similar to those we consider in our work. Partridge et. al. states in [8] that knowing which variables are the best in predicting user activities is important. Based on those essential variables, we have designed a model that arranges the generic properties of tasks and actions. The model represents three top-level ontologies: the task ontology, the action ontology and the SCP ontology, as shown in Fig. 1.

A task can be executed once, continuously or iteratively, which is captured in the *execution type* property. A task is also described by a *low-level context* property, which consists of *time* (day of week, hour of day) and *location*. The *location* property can be represented in two ways: coordinates (*latitude*, *longitude*) and a *symbolic location* (living room, kitchen, bedroom, and other spaces). The *user type* property describes information about *user status*, *age group*, *role* and *gender*. In addition, based on the number of users supported for a task, the *user interaction type* property represents single, shared and collaborative tasks. Controllable environmental-factors are represented in the environmental information dimension. The primitive properties like task name, synonym and URI are not shown in Fig.1 for simplicity. Partridge et. al. discovered that there are two essential combinations of variables that are effective to recognize user activities. The first combination is composed of the *location* and the *hour of the day* properties. Another combination is composed of the *age group* and the *day of the week* properties. We have reflected this result in our task selection algorithm by categorizing the properties into the primary – first combination – and secondary – second combination – groups.

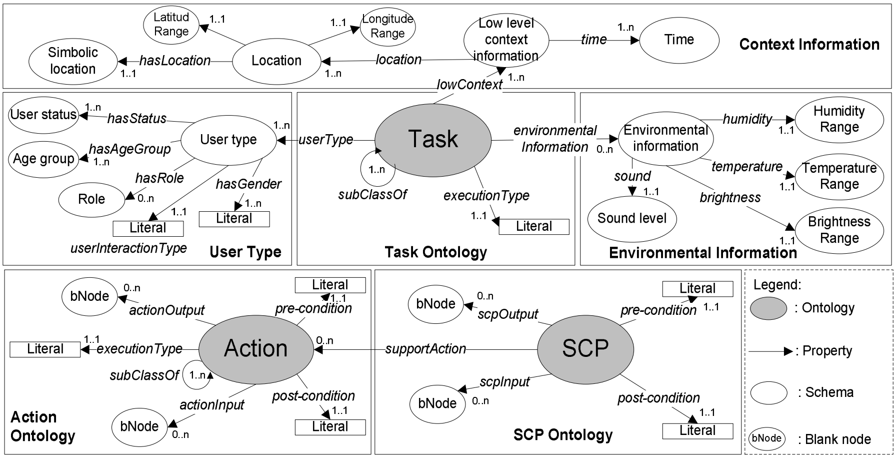


Fig. 1. Representation of Task, Action and SCP Semantics

Action and SCP properties. An action has to meet certain *pre-conditions* to be selected. In addition, execution of an action produces *post-conditions* that may affect the context of its environments. An action has *inputs* and *outputs*, which name and values are specified via a blank node (bNode). Primitive properties like action name, action URI and execution type are also hidden from the figure for simplicity. The action ontology also has all the properties in the context information, user type and environmental dimensions (these properties are not shown in Fig. 1 for simplicity).

3 Task Selection and Reconfiguration of Ubiquitous Applications

Task Selection Algorithm: Semantic Variables for Task Selection, and Matching of Properties. Let I be an input from the context manager, $I = (i_r, i_s)$, with i_r and i_s being primary and secondary variables respectively. Analogously, $P_T = (p_r, p_s)$, with

p_r and p_s being primary and secondary task properties respectively. In its turn, $p_r = \{p_{ld}, p_{wg}\}$, where p_{ld} represents the $location$ and $hour\ of\ day$ properties, and p_{wg} represents the $day\ of\ week$ and $age\ group$. Let p_s be the secondary task ontology properties. The task ontology hierarchy is composed by N nodes, where n denotes the n^{th} node. The total semantic matching value between I and P_T in relation to the node n, is defined as $V_n(I, P_T) = [V_R(i_r, p_r)_n \times W_R] + [V_S(i_s, p_s)_n \times W_S]$. $V_R(i_r, p_r)_n$ and $V_S(i_s, p_s)_n$ are the semantic matching values among the primary and secondary variables respectively. W_R and W_S are weights that reflects the significance of the variables. $W_R \geq W_S, W_R + W_S = 1$ and $W_R, W_S \in [0,1]$. In its turn, $V_R(i_r, p_r)_n = \sum_{r \in \{ld,wg\}} Match_n(i_r, p_r) \times w_r$, where $w_r \in \{w_{ld}, w_{wg}\}$ represents the weight of each primary variable. $Match_n(i_r, p_r)$ is equal to 1 if $Value(i_r) = Value_n(p_r)$, and 0 otherwise. $Value(i_r)$ is the primary input value and $Value_n(p_r)$ is the node n primary property value. When $Match_n(i_r, p_r) = 1, (\forall i_r \in I), (\forall p_r \in P_T)$, there is total semantic matching for the node n. All those nodes included in the set denoted by M are selected as candidates, whose secondary matching value $V_S(i_s, p_s)_m$ is then analogously computed, $m \in M$ (see Fig. 2).

On the contrary, if $\exists r \in \{ld, wg\}$ with $Match_n(i_r, p_r) = 0$, there is partial matching for the node n. Those nodes are proposed to the users for gathering their feedback of the selection. The total semantic matching value $V_m(I, P_T)$ for each $m \in M$ is obtained by weighting and then adding the primary and secondary values. Ranking those values the biggest one is selected, which corresponds to the selected task n^* (see Fig. 2). For nodes sharing the maximum value, they are compared based on their subsumption relationships, since the task ontology is hierarchically arranged based on tasks functionality. The mechanism continues with the SCP brokering process described in the next section, which shows how to get the $Actions(n^*)$ for the selected task n^*.

```
• Input: A set $I = (i_r, i_s)$, and a set $P_T = (p_r, p_s)$, output: A selected task $n^*$.
for each node $n \in N$ and $(\forall i_r \in I), (\forall p_r \in P_T)$ do
    $Match_n(i_r, p_r)$;
    $V_R(i_r, p_r)_n = \sum_{r \in \{ld,wg\}} Match_n(i_r, p_r) \times w_r$; end for
if there is not total matching do
    Users' feedback of partially matched nodes ; end if
else do
    Define the set $M$ of total matched nodes ;
    $V_S(i_s, p_s)_m$, $m \in M$;
    $V_m(I, P_T) = [V_R(i_r, p_r)_m \times W_R] + [V_S(i_s, p_s)_m \times W_S]$;
    Rank $V_m(I, P_T)$ ;
    Pick up the task $n^*$, where $V_{n^*}(I, P_T) = max\{V_m(I, P_T)\}$; end else
return $n^*$ ;
```

Fig. 2. The Task Selection Algorithm

Reconfiguration Algorithm: Semantic Variables, Matching of Properties and Semantic Distance of Actions. Let $Actions(n^*) = (a_1, a_2, \ldots, a_n)$ be the actions of a running task n^*, defined in the coordination model specified in the application template of n^*. $a_i \in A$ is an action of the action ontology A. Let i' be a change in any input from the context manager. Let p_{a_i} be the set of properties of an action a_i. The algorithm reported herein is denominated as a SCP brokering process**.** If $\exists a^* \in Actions(n^*), where\ Value(i') \neq Value(p_{a^*})$, a^* is reconfigured selecting a new action that fits well with the new input, by matching i' against p_{a_i} analogously to the matching process in the task selection algorithm.

Let Q be the set of total matched actions. Since the reconfiguration is done to preserve the semantics of the task, an equivalent action is selected to reconfigure the original action. Thus, a topological semantic distance of a^* is computed with all the $a_q \in Q$, which corresponds to the number of edges of the shortest path between them. The shorter the distance, the more similar the semantic is, since the action ontology is hierarchically arranged. The distances are ranked to select a_{q^*}, which has the smaller distance to a^*. In order to get an appropriate SCP from the SCP ontology, the node associated to a_{q^*} is selected, thus reconfiguring the task by $Actions(n^*, a_{q^*})$, i.e, a set of actions for task n^* including the new action a_{q^*} supported by its respective SCP. Finally, the list of abstract services of the SCP is used to perform a query into the service discovery, which is reported in [9]. The service discovery retrieves a list of available services, which are bound and executed according to the business process embedded into the application template.

4 Implementation and Evaluation

4.1 Implementation

The task selection and reconfiguration algorithms have been implemented for the service framework [10] for the Ubicomp middleware called Active Surroundings [11], [12]. The algorithms, as well as all the architectural elements of the framework, are implemented in Java, with the following support: (1) Protégé [13] has been used to create the task and action ontologies; (2) the Business Process Execution Language (BPEL) [14] is used to describe the coordination logic of an application template and a SCP; (3) the Jena library [15] has been used to implement the semantic reasoning algorithm to select tasks; (4) SPARQL [16] is used to define a query language and a protocol to access the ontology data represented in OWL [17] and (5) the ontology data and the Time-use Studies datasets are stored in a My-SQL [18] data base.

4.2 Experiments with Diverse Applications

The algorithms have proven to be effective on selecting and reconfiguring tasks by real demonstrations implemented in our test bed. An example of these demonstrations is the Ubiquitous Disk Jockey (UDJ) application. The goal of this demo is to make a Ubicomp environment to be appropriate for the people who dance in a party by playing music and controlling the room lights according to their levels of excitements. In order to realize the UDJ demo and other applications, the framework runs on a Windows XP platform on an Intel(R) Core(TM) 2, 1.87 GHz CPU with 2 GB of memory.

The test bed is equipped with a real-time location sensor system[1] that returns the coordinates of each user in the room. The test bed is also equipped to enable the context manager to generate input contextual information about brightness, temperature, humidity and several other contexts.

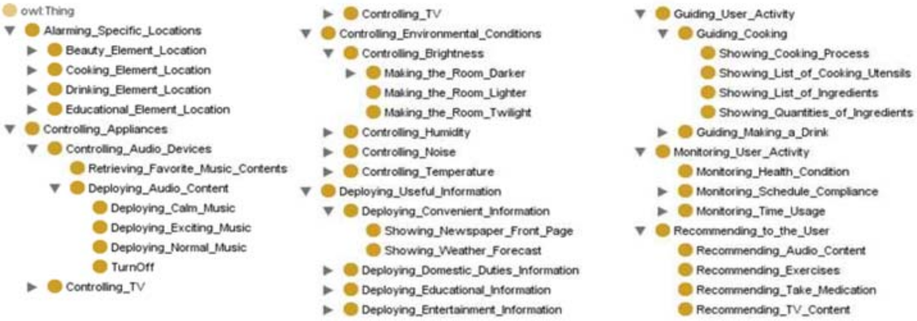
Each user in this demo carries a Wiimote to sense his/her motion, which is the controller for the Nintendo's Wii console. The context manager interprets each user's Wiimote raw motion data as a user's level of excitement. The context manager aggregates the individual data to recognize the level of excitement of a group of people, which is associated with the *users' status*. The result from the context manager includes the following primary context information and its values: (1) *symbolic location* (dancing area); (2) *hour of day* (night); (3) *day of week* (Saturday) and (4) *age group* (Young); and the secondary context information and its values: (1) *location* (coordinates of the dancing area); (2) *users' interaction type* (collaborative) and (3) *users' level of excitement* (calm). In this example, the value of the *users' interaction type* is "collaborative" because there are multiple people involved in the situation. The task selection algorithm returns the "dancing tasks" as the most appropriate task. By matching the above context information against the task properties and by narrowing down the matched tasks, the algorithm selects the task "dancing", in which the application template contains the actions "retrieving favorite music contents", "deploying calm music" and "making the room darker". According to the abstract services of the SCPs associated with those actions the service discovery finds the music search service, the audio service and the light control service. When users increase their motion while dancing, the level of excitement turns to "excited". When the users' level of excitement changes the task is reconfigured to play different music and to control lights to be appropriate to the new situation. The reconfiguration algorithm selects a new action by matching the new input with the action ontology properties, and measuring the semantic distance between the candidate actions and the action "deploying calm music". Then, the algorithm returns the new actions "deploying excited music" and "making the room lighter" for the task "dancing". In addition to the dancing party scenario explained above, we have produced several other scenarios and tested our approach with them (see Table 1 for some examples).

The major tasks defined for the Ubicomp home environment include metabolic, entertainment, sporting, personal care, duties, socializing, and other types of tasks. For instance, the task "having breakfast", a metabolic task, can be selected receiving

Table 1. Scenario Examples

Task	Actions
Having breakfast	Making the room lighter; deploying calm music; deploying useful information – *showing e-mail, wheatear forecast, e-newspapers front page, personal schedule.*
Watching TV	Deploying TV content; controlling TV content access; recommending TV content.
Making out	Making sport schedule; monitoring health condition; monitoring sport schedule compliance; recommending exercises; recommending medication.
Having a family party	Deploying music; deploying a background of family pictures; showing drinks mixing recommendations; guiding user for making drinks – *showing list of drinks, quantities of ingredients, recipes*; alarming specific locations – *alcohol and snack location.*

[1] Ubisense, based on ultra-wideband (UWB) technology, which delivers 15cm 3D positional accuracy of objects in real-time [19].



```
owl:Thing
▼ Alarming_Specific_Locations
    ▶ Beauty_Element_Location
    ▶ Cooking_Element_Location
    ▶ Drinking_Element_Location
    ▶ Educational_Element_Location
▼ Controlling_Appliances
    ▼ Controlling_Audio_Devices
          Retrieving_Favorite_Music_Contents
        ▼ Deploying_Audio_Content
              Deploying_Calm_Music
              Deploying_Exciting_Music
              Deploying_Normal_Music
              TurnOff
    ▶ Controlling_TV

    ▶ Controlling_TV
▼ Controlling_Environmental_Conditions
    ▼ Controlling_Brightness
        ▶ Making_the_Room_Darker
          Making_the_Room_Lighter
          Making_the_Room_Twilight
    ▶ Controlling_Humidity
    ▶ Controlling_Noise
    ▶ Controlling_Temperature
▼ Deploying_Useful_Information
    ▼ Deploying_Convenient_Information
          Showing_Newspaper_Front_Page
          Showing_Weather_Forecast
    ▶ Deploying_Domestic_Duties_Information
    ▶ Deploying_Educational_Information
    ▶ Deploying_Entertainment_Information

▼ Guiding_User_Activity
    ▼ Guiding_Cooking
          Showing_Cooking_Process
          Showing_List_of_Cooking_Utensils
          Showing_List_of_Ingredients
          Showing_Quantities_of_Ingredients
    ▶ Guiding_Making_a_Drink
▼ Monitoring_User_Activity
      Monitoring_Health_Condition
    ▶ Monitoring_Schedule_Compliance
    ▶ Monitoring_Time_Usage
▼ Recommending_to_the_User
      Recommending_Audio_Content
      Recommending_Exercises
      Recommending_Take_Medication
      Recommending_TV_Content
```

Fig. 3. Action Ontology: Different Actions to Support Tasks

the context information consisting on kitchen, morning, working day, adult, the coordinates of the kitchen table, and a single user's interaction type. Table 1 provides a detail of the actions included in the templates of those task scenarios. Fig. 3 shows a part of the action ontology developed to support those and other tasks. The model and algorithms that we developed are general enough to support diverse situations in the Ubicomp home environment.

4.3 Evaluation Results

The task selection algorithm was evaluated by running it on a Windows XP platform with an Intel Pentium 4 (3.33 GHz), and 896 MB of memory. The performance of this algorithm depends on the number of nodes in the task ontology. The algorithm was tested with different amounts of nodes: 200, 300, 400, 500, 600 and 700 nodes. The input data was randomly chosen from 65,635 records of users' tasks of the 2007 ATUS Time-use Study [7].

The algorithm is compared against the case without discrimination between primary and secondary properties. Fig. 4 shows that the time overhead is proportional to the number of nodes in the task ontology. In addition, this figure shows the effectiveness of prioritizing the properties of tasks. The result indicates that the prioritization

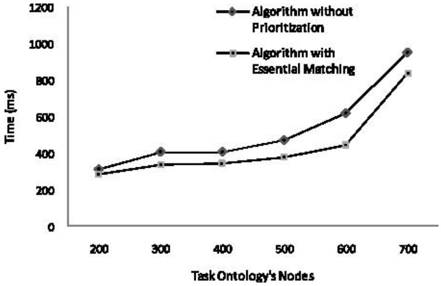


Fig. 4. Performance of the Task Selection Algorithm

of the properties contributes to decrease the execution time of the algorithm significantly. It is because the candidate nodes of the task ontology to consider can be narrowed down quicker by using the primary properties, matching the secondary variables with a smaller set of candidates.

5 Related Work

Although considerable research has been devoted to accurately and efficiently recognizes tasks based on context information [3], [4], [5], [6], less attention has been paid to the selection of coarse-grained tasks based on dynamic context, the reflection of users and system perspectives in different abstraction levels and the use of reusable composition patterns. The research has tended to focus on fine-grained tasks that do not properly realize user centricity. That is the case of the ABC framework, which provides a fine-grained tasks, services and data to dynamically adapt tasks to the available resources on heterogeneous computing devices [6]. In addition, some researches statically bind user's information against predefined applications. An example is the Gaia project that selects a new application by matching a user's data against a predefined application ontology [3]. Some research has focused on to provide solutions that directly selects and then integrates services that are equivalent to the user's task. An example is the IST AMIGO project, which uses predefined applications described as workflows that are directly matched against services, without any mediating process and abstract layers [5]. A similar example is the Aura project, which dynamically associates tasks to virtual services, but still there is a lack of an intermediate layer [4].

6 Conclusions and Future Work

In this paper, we tried to leverag the task-oriented computing approach to realize the user centricity required by Ubicomp environments. We have developed a semantically-based task and action description model in the Ubicomp home environment. This model is general enough to capture generic properties of tasks and actions. Additionally, this paper proposes a task selection and a task reconfiguration algorithm to bridge the gap between tasks and services. They are built on top of three composition layers: application templates, service composition patterns (SCPs) and services. The task selection algorithm implements a semantic and priority-based approach, and reacts to contextual information. The reconfiguration algorithm is designed to reflect the changes on context information for a task by using a pattern-based approach.

We are currently working on automatic service composition and reconfiguration to support spontaneous tasks in multi-spaces, especially in urban computing environments. We are also working on developing a task prediction mechanism based on of the Time-use Studies to extract users' behavioral patterns that can be used to prepare a task plan in advance.

Acknowledgments. This work was supported by the IT R&D program of MKE/IITA. [2008-F-047-02, Development of Urban Computing Middleware]. We thank to

Gonzalo Huerta-Canepa, Yong-Jae Lee, Wannasiri Bhuasiri, Kim Byoungoh, Jae-Sung Ku and SaeHyong Park.

References

1. Weiser, M.: The Computer for the 21st Century. In: Scientific American, pp. 94–104; Reprinted in IEEE Pervasive Computing, pp. 19–25. IEEE, Los Alamitos (2003)
2. Wang, Z., Garlan, D.: Task-Driven Computing. Technical Report, CMU - CS -00-154 (2000)
3. Roman, M., Hess, C.K., Cerqueira, R., Ranganathan, A., Campbell, R.H., Nahrstedt, K.: Gaia: A Middleware Infrastructure to Enable Active Spaces. In: IEEE Pervasive Computing, pp. 74–83. IEEE, Los Alamitos (2002)
4. Sousa, J.P., Garlan, D.: Aura: an Architectural Framework for User Mobility in Ubiquitous Computing Environments. In: 3rd Working IEEE/IFIP Conf on Software Architecture (2002)
5. Mokhtar, S.B., Liu, J., Georgantas, N., Issarny, V.: QoS-aware Dynamic Service Composition in Ambient Intelligence Environments. In: 20th IEEE/ACM International Conference on Automated Software Engineering, pp. 317–320. ACM Press, New York (2005)
6. Bardram, J.E.: From Desktop Task Management to Ubiquitous Activity Based Computing. In: Integrated Digital Work Environments, pp. 49–78. MIT Press, Cambridge (2007)
7. American Time-use Study, http://www.bls.gov/tus/
8. Partridge, K., Golle, P.: On Using Existing Time-Use Study Data for Ubiquitous Computing Applications. In: 10th Intl. Conference on Ubiquitous Computing, pp. 144–153. ACM Press, New York (2008)
9. Kang, S., Kim, W., Lee, D., Lee, Y.H.: Group Context-aware Service Discovery for Supporting Continuous Service Availability. In: 3rd International Workshop on Personalized Context Modeling and Management for UbiComp Applications (2005)
10. Jimenez-Molina, A.A., Koo, H., Ko, I.Y.: A Template-Based Mechanism for Dynamic Service Composition Based on Context Prediction in Ubicomp Applications. In: nternational Workshop on Intelligent Web Based Tools. IEEE ICTAI-07 (2007)
11. Huerta-Canepa, G.F., Jimenez-Molina, A.A., Ko, I.Y., Lee, D.: Adaptive Activity based Middleware. IEEE Pervasive Computing 7(2), 58–61 (2008)
12. Lee, D.: Active Surroundings: A Group-aware Middleware for Embedded Application Systems. In: 28th Intl. Conference on Computer Software and Applications, pp. 404–405. IEEE, Los Alamitos (2004)
13. The Protégé Ontology Editor and Knowledge-base Framework, http://protege.stanford.edu/
14. The Business Process Execution Language, http://docs.oasis-open.org/wsbpel/2.0/wsbpel-v2.0.pdf
15. The Jena Semantic Web Framework, http://jena.sourceforge.net/
16. The SPARQL query language for RDF, http://www.w3.org/2009/sparql/wiki/Main_Page
17. The OWL Web Ontology Language, http://www.w3.org/TR/owl-features/
18. The MySQL Data Base Management System, http://www.mysql.com/
19. Ubisense Real-time Location System, http://www.ubisense.net/

A Platform for Extracting and Storing Web Data

L. Víctor Rebolledo and Juan D. Velásquez

vireboll@gmail.com, jvelasqu@dii.uchile.cl

Abstract. Web data or data originated on the Web contain information and knowledge which allows to improve web site efficiency and effectiveness to attract and retain visitors.

However, web data have many irrelevant data inside. Consequently, it is necessary to preprocess them to model and understand the web user browsing behavior inside them. Further, due to frequent changes in the visitor's behavior, as well as in the web site itself, the discovered knowledge may become obsolete in a short period of time.

In this paper, we introduce a platform which extracts, preprocesses and stores web data to enabling the utilization of web mining techniques. In other words, there is an Information Repository (IR) which stores preprocessed web data and it facilitates the patterns extraction. Likewise, there is a Knowledge Base (KB) for storing the discovered patterns which have been validated by a domain expert.

The proposed structure was tested using a real web site to prove the effectiveness of our approach.

Keywords: Platform, Web mining, Knowledge Base.

1 Introduction

Web data let extract information and knowledge that suggests changes to become a web site more effective and efficient [8][11]. This potential is determined by a site's content, its design and structure [6]. Indeed, through web log analysis is possible to understand the visitor's behavior, and further, together with web page content processing, it is possible extract which the visitor's content preferences are [9][10].

Nevertheless, web data have to be cleaned, consolidated and transformed in an ad-hoc structure for the application of web mining [1][8]. Particularly, web logs have many irrelevant data that should be cleaned; web site text content should be processed by removing HTML tags and by taking it to an appropriate structure, for example *Vector Space Model* [5]. As a result, it implies high costs in time and resources.

In addition, the visitor's interests change in the time, as well as in the web site itself. Furthermore, the discovered knowledge may become obsolete in a short period of time [8]. In other words, we periodically have to incur in the cost of preprocess web data to analyze the web site.

The proposed platform allows keeping web data ready to apply web mining. On that point, we can extract knowledge whenever we want and without incur in costs associated to preprocess web data.

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 838–845, 2009.

2 The Platform

The proposed platform is formed by different components such as:

- **Web site**: composed for web logs, web page text content and other objects like images, files, etc[1].
- **Data Staging Area or DSA**: It is the area where web data are cleaned and preprocessed. For this paper, we use a relational database as DSA.
- **Extraction Process**: Through it, the data are periodically extracted from web site content and web logs to the DSA
- **Information Repository**: it is a repository, built with Data Warehouse architecture[2], where the preprocessed web data are stored. It facilitates the extraction of feature vectors which are the input for web mining algorithms.
- **Transformation and Load Process**: Corresponds to sessionization process (see subsection 3.1) and transformation of web site text content into a Vector Space Model (see subsection 3.2). The transformation happen in DSA and the data are loaded in Information Repository.

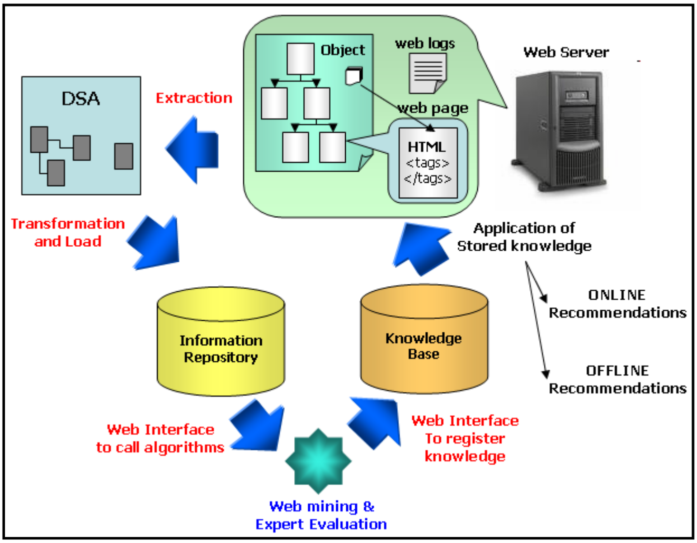


Fig. 1. Platform components

- **Web interface to generate vectors**: Through them, the data miners can create specific vectors for their web mining studies.
- **Web mining algorithms**: Corresponds to the techniques used to extract knowledge from web data (see subsection 3.5)

[1] In order to assure web data provision, we use Plone, a Content Management System, to manage the web site content. Indeed, this tool let add metadata to each web object and keep records about web site's changes.

- **Knowledge Base KB**: It is a repository, as Information Repository, where the discovered knowledge will be stored.
- **Web interface to register knowledge**: The extraction of knowledge is not an automatic process because it requires the domain expert's interpretation and validation [7]. For this reason, this web interface will let data miners manually register the discovered knowledge in KB (see subsection 4.4).

3 Modeling and Preprocessing Web Data

In order to apply web mining techniques, web logs and web pages have to be preprocessed by using specific models to representing them.

3.1 Preprocessing of Web Logs

For each visitor, it is necessary to determine the sequence of pages during a session based on web logs entries. This process is called **sessionization[9]** and it can be performed by using tables and program filters. We consider a maximum time duration of 30 minutes per session and we use only web logs registers with non-errors codes chose URL parameters link to web page objects.

3.2 Preprocessing of Web Site

A good representation to web site text content is the Vector Space Model[5]. Indeed, let R be the number of different words in a web site and Q the number of web pages. A vectorial representation of the web site is a matrix M of dimension RxQ, $M = (m_{ij})$ where $i = 1, \dots, R, j = 1, \dots, Q$ and m_{ij} is the weight of the *i*th word in the *j*th page. To calculate these weights, we use a variant of the *tfxidf-weighting[5]*, defined as follows:

$$m_{ij} = f_{ij}\left(1 + sw(i)\right) * log\left(\frac{Q}{n_i}\right) \tag{1}$$

Where f_{ij} is the number of occurrences of the *i*th word in the *j*th page, *sw(i)* is a factor to increase the importance of special words and n_i is the number of documents containing the *i*th word. A word is special if it shows special characteristics, e.g., the visitor searches for this word.

Definition 1 (Page Vector). It is a vector $WP^j = \left(wp_1^j, \dots, wp_R^j\right) = (m_{1j}, \dots, m_{Rj})$ with $j = 1, \dots,$ Q, that represent a list of words contained within a web page. It represents the jth page by the weights of the words contained in it, i.e., by the jth column of M. The angle's cosine is used as a similarity measure between two page vectors:

$$dp(WP^i, WP^j) = \frac{\sum_{k=1}^{R} wp_k^i wp_k^j}{\sqrt{\sum_{k=1}^{R}\left(wp_k^i\right)^2}\sqrt{\sum_{k=1}^{R}\left(wp_k^j\right)^2}} \tag{2}$$

3.3 Modeling the User Browsing Behavior

Our visitor behavior model uses three variables: the sequence of visited pages, their contents and the time spent on each page. The model is based on a *n*-dimensional visitor behavior vector which is defined as follows.

Definition 2 (User Behavior Vector UBV). It is a vector $v = [(p_1, t_1), \dots, (p_n, \dots, t_n)]$, where the pair$(p_i, t_i)$ represents the *i*th page visited p_i and the percentage of time spent on it within a session t_i, respectively.

3.3.1 Comparing User Behavior Vectors

Let α and β be two visitor behavior vectors of dimension C^α and C^β, respectively. Let $\Gamma(.)$ be a function that returns the navigation sequence corresponding to a visitor vector. A similarity measure has been proposed elsewhere to compare visitor sessions, as follows [9]:

$$sm(\alpha, \beta) = dG(\Gamma(\alpha), \Gamma(\beta)) \frac{1}{\eta} \sum_{k=1}^{\eta} \tau_k * dp(p_{\alpha,k}, p_{\beta.k}) \tag{3}$$

where $\eta = \min\{\mathrm{C}^\alpha, \mathrm{C}^\beta\}$, and $dp(p_{\alpha,k}, p_{\beta,k})$ is the similarity between the *k*th page of vector α and the *k*th page of vector β. The term $\tau_k = min\{t_{\alpha,k}/t_{\beta,k}, t_{\beta,k}/t_{\alpha,k}\}$ is an indicator of the visitor's interest in the visited pages. The term *d*G is the similarity between sequences of pages visited by two visitors [9].

3.4 Modeling the User's Text Preferences

A *web site keyword* is defined as a word or a set of words that makes the web page more attractive to the visitor [10]. The task here is to identify which are the most important words (keywords) in a web site from the visitor's viewpoint. This is done by combining usage information with the web page content and by analyzing the visitor behavior in the web site.

To select the most important pages, it is assumed that the degree of importance is correlated with the percentage of time spent on each page within a session. By sorting the visitor behavior vector according to the percentage of time spent on each page, the first ι pages will correspond to the ι -most important pages.

Definition 3 (ι-most Important Page Vector IPV): It is a vector $\vartheta(v)[(\rho_1, t_1), \dots, (\rho_\iota, t_\iota)]$, where the pair (ρ_ι, t_ι) represents the ιth most important page and the percentage of time spent on it within a session.

3.4.1 Comparing Important Page Vector

Let α and β be two visitor behavior vectors. A similarity measure between two ι most important pages vectors is defined as:

$$st(\vartheta_\iota(\alpha), \vartheta_\iota(\beta)) = \frac{1}{\eta} \sum_{k=1}^{\iota} min \left\{ \frac{\tau_k^\alpha}{\tau_k^\beta}, \frac{\tau_k^\beta}{\tau_k^\alpha} \right\} * dp(\rho_k^\alpha, \rho_k^\beta) \tag{4}$$

where the term $min\{.,.\}$ indicates the visitors' interest in the visited pages, and the term *dp* is the similarity measure (2)

In (4), the similarity of the most important pages is multiplied by the ratio of the percentage of time spent on each page by visitors α and β. This allows us to distinguish between pages with similar contents, but corresponding to different visitors' interests.

3.5 Applying Web Mining Techniques

Due to most of times the visitors are anonymous, there is no previous idea about visitor behavior, and hence clustering techniques are useful [7][10]. In that sense, we use 2 clustering algorithms to validate the obtained patterns. Moreover, we extract association rules to find correlations between the pages visited into a session.

- **Identifying Association Rules:** By using the classic algorithm Apriori[4], we can validate or reject the patterns obtained with another technique like clustering

- **Clustering UBV:** We apply Self Organizing Feature Maps SOFM and K-means on UBV by using the similarity measure (3). Firstly, we use SOFM which requires vectors of the same dimension. Let H be the dimension of the UBV. If a user session has less than H elements, the missing components up to H are filled with zeroes. Else if the number of elements is greater than H only the first H components are considered. Later, we use K-means algorithm by setting the number of clusters as the amount of validated clusters obtained with SOFM.

- **Clustering IPV:** A SOFM y K-means are used to find groups of similar user sessions. The most important words for each cluster are determined by identifying the cluster centroids. The importance of each word according to each cluster is calculated by:

$$kw[i] = \sqrt[\iota]{\prod_{p\in\zeta} m_{ip}} \tag{5}$$

for $i = 1, \dots, R$, where *kw* is an array containing the geometric mean of the weights of each word (1) within the pages contained in a given cluster. Here, ζ is the set of pages contained in the cluster. By sorting *kw* in descending order, the most important words for each cluster can be selected.

4 Real-World Application

The above described methodology was applied to the web site of University of Chile's Web Intelligence Research Group (http://wi.dii.uchile.cl). This site was built by using Plone and it is formed by 42 static web pages in Spanish and English, and by 102 objects as files and pictures. We analyzed 72.481 all the visits done in the period from January to May, 2008. Approximately, 80 thousands of raw log registers were collected.

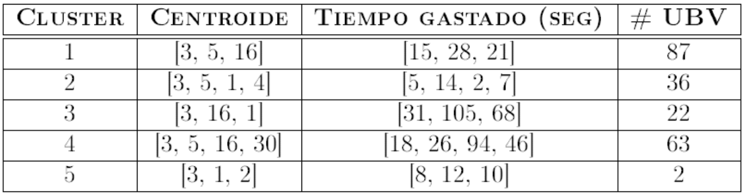
4.1 Knowledge Extracted from Association Rules

Due to we identify a little amount of sessions, we need to adjust the confidence and support levels in order to get interesting association rules. With a confidence level of 16%, we found high correlation between pages: *Bienvenido (Welcome), Estudiantes (Students), Investigación (Research)* and *Investigadores (Researchers).* Indeed, three of them appear together in 16,1% of sessions.

4.2 Knowledge Extracted from Visitor Browsing

After applying SOFM to UBV, five clusters were found. These clusters are presented in **Table 1**. Indeed, the second column of this table contains the centroid (winner neuron) of each cluster, representing the sequence of the pages visited. The third column contains the time spent in each page and the fourth column, the amount of vectors that belongs to each clusters. The last centroid only represents 2 vectors, furthermore, we only considerate four relevant clusters.

Table 1. Clusters found with SOFM



Cluster	Centroide	Tiempo gastado (seg)	# UBV
1	[3, 5, 16]	[15, 28, 21]	87
2	[3, 5, 1, 4]	[5, 14, 2, 7]	36
3	[3, 16, 1]	[31, 105, 68]	22
4	[3, 5, 16, 30]	[18, 26, 94, 46]	63
5	[3, 1, 2]	[8, 12, 10]	2

In that sense, we set K-means to obtain the following four clusters:

- **Cluster 1**: Visitors which were searching information about the members' publications of WI Group.
- **Cluster 2**: Visitors which were searching academic information about the members of WI Group.
- **Cluster 3**: Visitors which were interested on activities of teaching and seminars imparted by WI Group
- **Cluster 4**: Visitors which were searching information about the studies made by WI Group which are of public interest

4.3 Knowledge Extracted from Visitor Preferences

After applying the SOFM to the 3-most important pages vectors, five clusters were found, however only four clusters were considered. These clusters can be seen in **Fig. 2**. Applying (5), we obtained the keywords and their relative importance in each cluster. For example, the cluster 1 = {3, 16, 1}, and $kw[i] = \sqrt[3]{m_{i3}m_{i16}m_{i1}}$ with $i = 1, \dots, R$. By sorting kw[i], the group of most important words for each cluster were selected. Some of the keywords found were: *web, development, technology, data, professor, graduated, information, base*, etc.

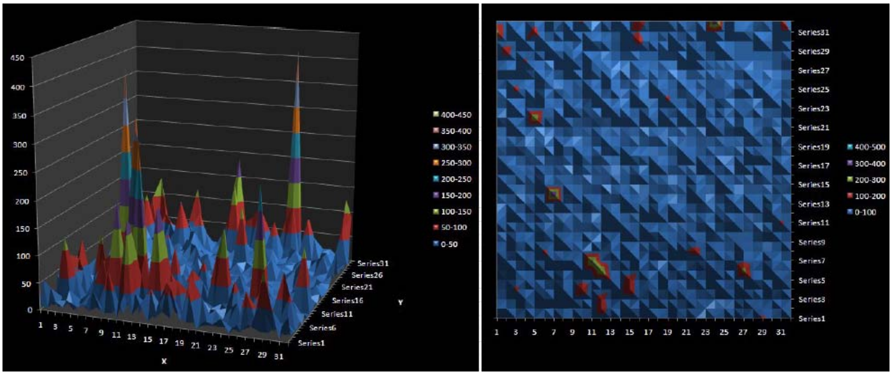


Fig. 2. Identifying clusters by using SOFM

At the same way and by using the above IPV, we apply K-means to get four clusters (we set it a priori). Some of the keywords found were: *information, investigation, title, graduated, professor, system, courses, knowledge, intelligent*, etc.

4.4 Loading the Knowledge Base

Through a web interface, we can introduce the patterns interpreted with the expert help. The Knowledge base will be formed by a Fact Table with the studies' results and Dimension tables with attributes which characterize the discovered patterns. In that sense, the KB stores the web mining technique used (WMT), the date when the technique was applied, the found pattern and its interpretation, etc. For example:

- **Time:** (2008, July, 23, 05:30 hrs.)
- **Browsing_Behavior:** The clusters centroids discovered by the mining process.
- **WMT:** SOFM with thoroidal neighbor and 32*32 neurons
- **Text_Preference:** The keywords context of the discovered clusters.

This information can be used by a human for changes in web site structure and content and by a system which makes navigation recommendations to the user when they present a behavior pattern that coincide with some stored in the repository [8].

5 Conclusions

The proposed platform preprocesses web data periodically in order to always have available vectors to enter to web mining algorithms. Moreover it allows storing discovered patterns in a repository, which can be used by a computational system through a set of rules that make online navigation recommendations and by humans for offline changes in the web site content and structure. Consequently, the web site can be modified in order to make it more efficient and effective to attract and retain users. In future works, other web mining techniques will be applied in order to provide new patterns and rules for the KB.

Acknowledgements

The authors thank to the Millennium Scientific Nucleus on Complex Engineering Systems (Chile), which has partially funded this work.

References

[1] Cooley, R.W.: Web usage mining: discovery and application of interesting patterns from web data, Dissertation for degree of Doctor of Philosophy. University of Minnesota, Faculty of the Graduate School, Minnesota, USA (2000)

[2] Kimball, R., Merx, R.: The Data Webhouse Toolkit. Wiley Computer Publisher, Chichester (2000)

[3] Kosala, R., Blockell, H.: Web mining research: a survey. SIGKDD Explorations: Newsletter of the Special Interest Group (SIG) on Knowledge Discovery and Data Mining 2(1), 1–15 (2000)

[4] Larose, D.T.: Discovering Knowledge in Data: An Introduction to Data Mining. John Wiley & Sons, Chichester (2005)

[5] Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Communications of the ACM 18(11), 613–620 (1975)

[6] Srivastava, J., Cooley, R., Deshpande, M., Tan, P.: Web usage mining: discovery and applications of usage patterns from web data. SIGKDD Explorations 1(2), 12–23 (2000)

[7] Theodoridis, S., Koutroumbas, K.: Pattern Recognition. Academic Press, London (1999)

[8] Velasquez, J.D., Palade, V.: Adaptive Web site: A Knowledge Extraction from Web Data Approach. IOS Press, Amsterdam (2008)

[9] Velasquez, J.D., Yasuda, H., Aoki, T., Weber, R.: A new similarity measure to understand visitor behavior in a web site. IEICE Transactions on Information and Systems, Special Issues in Information Processing Technology for web utilization E87-D(2), 389–396 (2004)

[10] Velasquez, J.D., Yasuda, H., Aoki, T., Weber, R., Vera, E.: Using self organizing feature maps to acquire knowledge about visitor behavior in a web site. In: Palade, V., Howlett, R.J., Jain, L. (eds.) KES 2003. LNCS (LNAI), vol. 2773, pp. 951–958. Springer, Heidelberg (2003)

[11] Yao, Y.Y.: Web intelligence: New frontiers of exploration. In: Proceedings of the 2005 International Conference on Active Media Technology (AMT 2005), Takamatsu, Kagawa, Japan, May 19-21 2005, pp. 3–8 (2005)

Bayesian Reflectance Component Separation

Ramón Moreno, Manuel Graña*, Alicia d'Anjou, and Carmen Hernandez

Grupo de Inteligencia computacional,
University of the Basque Country
www.ehu.es/ccwintco

Abstract. We work on a Bayesian approach to the estimation of the specular component of a color image, based on the Dichromatic Reflection Model (DRM). The separation of diffuse and specular components is important for color image segmentation, to allow the segmentation algorithms to work on the best estimation of the reflectance of the scene. In this work we postulate a prior and likelihood energies that model the reflectance estimation process. Minimization of the posterior energy gives the desired reflectance estimation. The approach includes the illumination color normalization and the computation of a specular free image to test the pure diffuse reflection hypothesis.

1 Introduction

Works on reflectance map estimation [17,4,10,9,12,1,2] usually need to impose some assumptions like the knowledge of a color segmentation of the image, the detection of color region boundaries or color discontinuities, or the knowledge of the decomposition into linear basis functions of the surface color. The approach presented here does not impose any such assumption and does not need previous segmentations of the image. Most of the works in the literature are based on the Dichromatic Reflection Model (DRM) [8], and we will also follow this model for the development of our approach. We follow a Bayesian approach [3] to model the desired result as constraints implemented in an *a priori* distribution. We postulate the *a priori* distribution based on the idea developed in [13] that the derivatives of the logarithmic images of both diffuse image and specular free must be equal in order to have pure diffuse pixels.

Section 2 gives the reflection modelling background, section 3 describes our Bayesian model giving the expressions for the *a priori* and likelihood energies. Section 4 presents some experimental results. Section 5 gives some summary conclusions and ideas for further work.

2 Reflection Modelling

The DRM was proposed by Shafer [8]. It describes the surface reflection of light in dielectric materials as the sum of two components, the diffuse and specular

* The Spanish Ministerio de Educacion y Ciencia supports this work through grant DPI2006-15346-C03-03.

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 846–852, 2009.

terms. The diffuse reflection component exhibits the color of the material due to different light wavelengths are more or less absorbed as light is scattered by the material. The specular reflection component is essentially determined by the color of incident light. The model of the image taken with a digital camera is as follows

$$\mathbf{I}(x) = w_d(x) \int_{\Omega} S(\lambda, x) E(\lambda) \mathbf{q}(\lambda) d\lambda + w_s(x) \int_{\Omega} E(\lambda) \mathbf{q}(\lambda) d\lambda \qquad (1)$$

$$I(x) = w_d(x)\mathbf{B}(x) + w_s(x)\mathbf{G}, \qquad (2)$$

where $\mathbf{I} = \{I_r, I_g, I_b\}$ is the color of an image pixel obtained through a camera sensor, $x = \{x, y\}$ are the two dimensional coordinates of the pixel in the image, $\mathbf{q} = \{q_r, q_g, q_b\}$ is the three element vector of sensor sensitivity, and $w_d(x)$ and $w_s(x)$ are the weighting factors for diffuse and specular components, respectively, which depend on the geometric structure at location x, $S(\lambda, x)$ is the diffuse spectral reflectance, $E(\lambda)$ is the illumination spectral power distribution function, which is independent of the spatial location x because we assume a uniform illumination color. The integration is done over the visible light spectrum Ω. We define the following chromatic terms:

- Image chromaticity (normalized RGB space) : $\mathbf{\Psi}(x) = \frac{\mathbf{I}(x)}{I_r(x)+I_g(x)+I_b(x)}$
- Diffuse Chromaticity: $\mathbf{\Lambda}(x) = \frac{\mathbf{B}(x)}{B_r+B_g+B_b}$
- Specular or Illumination Source Chromaticity: $\mathbf{\Gamma} = \frac{\mathbf{G}}{G_r+G_g+G_b}$.

The image can be written in terms of diffuse an specular chromaticity $\mathbf{I}(x) = m_d(x)\mathbf{\Lambda}(x) + m_s(x)\mathbf{\Gamma}$, where $m_d(x) = w_d(x)\left[B_r(x) + B_g(x) + B_b(x)\right]$ and $m_s(x) = w_s(x)\left(G_r + G_g + G_b\right)$. We can see that the diffuse chromaticity depends on the pixel location x, while the specular chromaticity does not, because we assume a uniform illumination color. Both weighting factors depend on the geometric structure at location x.

For the ensuing processes, we will be assuming that illumination colors (the specular component) will be pure white, so that $\Gamma_r = \Gamma_g = \Gamma_b$. The illumination corrected image is computed as $\mathbf{I}'(x) = \frac{\mathbf{I}(x)}{\mathbf{\Gamma}^{est}(x)}$. Where Γ^{est} is the estimation of the illumination color, that can be estimated by some of the methods proposed in the literature [6,14,15,2]. The quotient is computed as the Hadamard quotient (applied at each component independently). The normalized image can be expressed as $\mathbf{I}'(x) = m'_d(x)\mathbf{\Lambda}'(x) + \frac{m'_s(x)}{3}$, where $\mathbf{\Lambda}'$ is the illumination color normalized diffuse chromaticity.

It is possible to obtain an specular free image [17] from the color normalized image by the following procedure:

1. Compute at each pixel the minimum of all of its three color bands $\tilde{\mathbf{I}}(x) = min\{I'_r(x), I'_g(x), I'_b(x)\}$, therefore $\tilde{\mathbf{I}}(x) = m'_d(x)\tilde{\mathbf{\Lambda}}(x) + \frac{m'_s(x)}{3}$, where $\tilde{\mathbf{\Lambda}}(x) = min\{\Lambda'_r(x), \Lambda'_g(x), \Lambda'_b(x)\}$.

2. Compute at each pixel the difference of the normalized image and the one obtained in the previous step $\mathbf{I}^{sf}(x) = \mathbf{I}'(x) - \tilde{\mathbf{I}}(x) = m'_d(x)\left[\boldsymbol{\Lambda}'(x) - \tilde{\boldsymbol{\Lambda}}(x)\right]$, so that the specular component dissapears from the image.

2.1 Separation Method

We will base our Bayesian model in the key element of the method proposed in [11,13]. The pure diffuse pixels can be characterized by the following relation:

$$\triangle(x) = dlog(\mathbf{I}'(x)) - dlog(\mathbf{I}^{sf}(x)) = 0, \tag{3}$$

where $dlog(\mathbf{I}^{sf}(x)) = \frac{\partial}{\partial x} log(\mathbf{I}^{sf}(x))$ and $dlog(\mathbf{I}'(x)) = \frac{\partial}{\partial x} log(\mathbf{I}'(x))$, the logarithm is computed pixel wise, and the spatial derivative can be computed in several ways, for instance in [13] it is computed on the scalar value image given by the summation of the three channels. It can be easily verified that $dlog(\mathbf{I}'(x)) = \frac{\partial}{\partial x} log(m'_d(x)) = dlog(\mathbf{I}^{sf}(x))$ for pure diffuse pixels if the diffuse chromaticity of neighboring pixels is the same. That means that the method works well inside homogenous color regions, and needs the estimation of color region boundaries. When $\triangle(x) > 0$ in eq. 3 and the pixel is not at a color boundary and a pure specular pixel, then it has some specular component that can be removed to get the diffuse reflectance component. The method proposed in [13] follows from an heuristic observation about the distribution of pixels in the maximum chromaticity versus (normalized illumination color) intensity space. Non diffuse pixels are decreased in intensity iteratively to search for the pure diffuse pixel value.

3 Bayesian Modelling

Given an image f and a desired unknown response of a computational process d, Bayesian reasoning gives, as the estimate of d, the image wich maximizes the *A Posteriori* distribution $P(d|f) \propto e^{-U(d|f)}$, where the *A Posteriori* energy can be decomposed in to the *A Priori* $U(d)$ and Likelihood (Conditional) $U(f|d)$ energies $U(d|f) = U(f|d) + U(d)$. The Maximum A Posteriori (MAP) estimate is equivalent minimize the posterior energy function

$$d^* = \arg\min_d U(d|f) \tag{4}$$

The Likelihood energy $U(f|d)$ measures the cost caused by the discrepancy between the input image f and the solution d. The A Priori energy $U(d)$ is a model of the desired solution, usually built as a Random Markov Field (RMF), so that the A Priori energy can be built up as the summation of the local energies at the pixels, which are expressed as summations over the set of cliques including the pixel, weighted by the local potential parameter. A Priori energy usually incorporates any desired constraint, such as smoothness, into the model.

We will assume a Gaussian Likelihood distribution plus a Chromaticity preservation constraint, therefore the Likelihood energy will have the following expression:

$$U(d|f) = \sum_{i=1}^{m} \frac{(f_i - d_i)^2}{2\sigma^2} + \sum_{i=1}^{m} \left(\Psi_i^f - \Psi_i^d \right)^2 ,$$

where f_i and d_i are the RGB pixel values a the i-th pixel position for the observed and desired image, respectively. Also, Ψ_i^f and Ψ_i^d denote the chromaticity pixels of the observed and desired image, respectively.

The A Priori energy is built up from two components. The first one is the Chromaticity continuity:

$$U_{\Psi}(d) = \sum_{i=1}^{m} \sum_{j \in N_i} \sum_{c \in \{r,g,b\}} \left(\Psi_{i,c}^d - \Psi_{j,c}^d \right)^2 .$$

The second modelling the estimation of the derivatives in eq. 3 as the cliques of the RMF. That is, we assume that the local energy at pixel d_i is defined as

$$U_{\triangle}\left(d_i\right) = \left(dlog(d_i) - dlog(d_i^{sf}) \right)^2 ,$$

where d_i^{sf} is the i-th pixel of the specular free image, computed as described above, and $dlog(.)$ in means the local estimation of the derivative, which is approximated as follows:

$$dlog(d_i) = \frac{1}{\#N} \sum_{j \in N_i} \log(\frac{I(x_j)}{I(x_i)}),$$

where N_i is the local neighborhood of pixel d_i, and $\#N$ is its cardinality. After some manipulations, the local derivative component of the A Priori energy is derived as:

$$U_{\triangle}\left(d_i\right) = \left(\sum_{j \in N_i} \sum_{c \in \{r,g,b\}} \log \frac{d_{j,c} d_{i,c}^{sf}}{d_{i,c} d_{j,c}^{sf}} \right)^2 .$$

This local energy is equivalent to the Kuk-Jin ratio criterion [17]. The derivative component of the A Priori energy is, therefore, the addition of these local energies:

$$U_{\triangle}\left(d\right) = \sum_{i=1}^{m} U\left(d_i\right),$$

and the A Priori energy is given by the addition $U\left(d\right) = U_{\triangle}\left(d\right) + U_{\Psi}\left(d\right)$.

4 Some Experimental Results

In this section we report some experimental results applying the Bayesian approach described above. The starting value for the energy minimization process

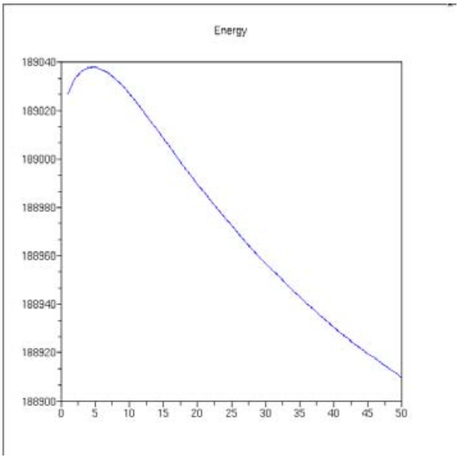


Fig. 1. Evolution of the energy function in an instance run of the algorithm

is set to $f = d(0) = \mathbf{I}'$. Each iteration step of the energy minization involves the computation of the specular free image $d^{sf}(t)$ of the current hypothesis $d(t)$ of the optimal estimation d^*. Instead of using a Monte Carlo minimization technique [3], such as Simulated Annealing, we have employed a simple heuristic to determine the new hypothesis $d(t+1)$, consisting in the reduction of the intensity of the pixels preserving their chromacity components relative ratios. Although simple, this strategy does in fact produce a minimization of the energy function, as can be appreciated in figure 1, where we plot an instance of the energy function evolution. We have tested our approach on some images already tested by some authors in the literature i.e. [13,12] among others. Figure 2 shows the result over a well known test image with two colors and two light sources. Our algorithm does not include any modelling of the underlying color regions in the scene, such as in [12], so it can be appreciated that the almost pure specular pixels can not be corrected, because there almost no chromatic information left in them. To improve our approach we will be including a color map field in the model, to be able to assign those pixels the most likely color. The figure 3 shows a complex geometry image. Our estimation of the diffuse reflectance component recovers the underlying geometry, with some blurring effects.

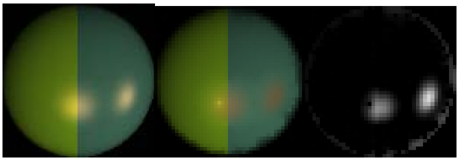


Fig. 2. From left to righ, the original image, the estimated diffuse reflection component, and the estimated especular component

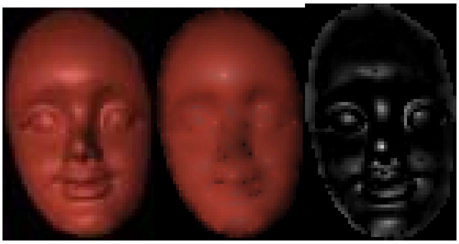


Fig. 3. From left to righ, the original image, the estimated diffuse reflection component, and the estimated especular component

5 Conclusions and Further Works

We have presented a Bayesian approach to the problem of reflection component separation. As in previous works, our approach works with only one image [13] and does not need any additional assumption, such as models of the colors in scene o previous color segmentations of the image. We compute the specular free image, which can be done on the fly for each hypothesis. We have tested the approach applying a simple heuristic to provide new hypothesis from the previous iteration, with quite encouraging results. From the experiments we detect the need to incorporate a color map field in the *A Priori* model, so that the color of almost purely specular pixels can be recovered more easily. The problem of diverse color illumination sources will be dealt with in further works. We will also extend our works to other imaging models [7,16,5].

References

1. Ma, W.-C., Hawkins, T., Peers, P., Chabert, C.-F., Weiss, M., Debevec, P.: Rapid acquisition of specular and diffuse normal maps from polarized spherical gradient illumination. In: Eurographics Symposium on Rendering 2007 (2007)
2. Fu, Z., Tan, R.T., Caelli, T.: Specular free spectral imaging using orthogonal subspace projection. In: 18th International Conference on Pattern Recognition, 2006. ICPR 2006, vol. 1, pp. 812–815 (2006)
3. Winkler, G.: Image analysis, random fields and dynamic Monte Carlo methods. Springer, Heidelberg (1995)
4. Hara, K., Nishino, K., Ikeuchi, K.: Light source position and reflectance estimation from a single view without the distant illumination assumption. IEEE Trans. Pattern Anal. Mach. Intell. 27(4), 493–505 (2005)
5. Jensen, H.W., Marschner, S.R., Levoy, M., Hanrahan, P.: A practical model for subsurface light transport. In: Proceedings of the 28th annual conference on Computer graphics and interactive techniques, pp. 511–518. ACM Press, New York (2001)
6. Choi, Y.-J., Yoon, K.-J., Kweon, I.S.: Illuminant chromaticity estimation using dichromatic slope and dichromatic line space. In: Korea-Japan Joint Workshop on Frontiers of Computer Vision, pp. 219–224. FCV (2005)

7. Phong, B.T.: Illumination for computer-generated images. PhD thesis, The University of Utah (1973)
8. Shafer, S.A.: Using color to separate reflection components. Color Research and Aplications 10, 43–51 (1984)
9. Tan, R.T., Nishino, K., Ikeuchi, K.: Color constancy through inverse-intensity chromaticity space. J. Opt. Soc. Am. A Opt. Image Sci. Vis. 21(3), 321–334 (2004)
10. Tan, R.T., Nishino, K., Ikeuchi, K.: Separating reflection components based on chromaticity and noise analysis. IEEE Trans. Pattern Anal. Mach. Intell. 26(10), 1373–1379 (2004)
11. Tan, R.T., Ikeuchi, K.: Separating reflection components of textured surfaces using a single image. In: Proceedings of Ninth IEEE International Conference on Computer Vision, 2003, October 13-16, 2003, vol. 2, pp. 870–877 (2003)
12. Tan, R.T., Ikeuchi, K.: Reflection components decomposition of textured surfaces using linear basis functions. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005, June 20-25, 2005, vol. 1, pp. 125–131 (2005)
13. Tan, R.T., Ikeuchi, K.: Separating reflection components of textured surfaces using a single image. IEEE Transactions on Pattern Analysis and Machine Intelligence 25(2), 178–193 (2005)
14. Toro, J.: Dichromatic illumination estimation without pre-segmentation. Pattern Recogn. Lett. 29, 871–877 (2008)
15. Toro, J., Funt, B.: A multilinear constraint on dichromatic planes for illumination estimation. IEEE transactions on image processing: a publication of the IEEE Signal Processing Society 16, 92–97 (2007) PMID: 17283768
16. Ward, G.J.: Measuring and modeling anisotropic reflection. SIGGRAPH Comput. Graph. 26, 265–272 (1992)
17. Yoon, K.-J., Choi, Y., Kweon, I.S.: Fast separation of reflection components using a specularity-invariant image representation. In: 2006 IEEE International Conference on Image Processing, October 8-11, 2006, pp. 973–976 (2006)

Identifying Fewer Key Factors by Attribute Selection Methodologies to Understand the Hospital Admission Prediction Pattern with Ant Miner and C4.5

Kyoko Fukuda

Geo Health Lab, Department of Geography, University of Canterbury, Private Bag 4800,
Christchurch, New Zealand
kyoko.fukuda@canterbury.ac.nz

Abstract. Attribute Selection (AS) is generally applied as a data pre-processing step to sufficiently reduce the number of attributes in a dataset. This study uses six different data mining AS methods to identify a few key driving climate and air pollution attributes from small attribute sets (16 attributes) to increase knowledge about the underlying structures of acute respiratory hospital admission counts, because understanding key factors in environmental science data helps constructing a cost effective data collection and management process by focusing on collecting and investigating more representative and important variables. The performance of the selected attribute set was tested with Ant-Miner and C4.5 classifiers to examine the ability to prediction the admission count. Removal of attributes was successful over all AS methods, especially TNSU (a newly developed AS method, Tree Node Selection for unpruned), which achieved best in removing attributes and some improving the classification accuracy for Ant-Miner and C4.5. However, the overall prediction accuracy improvements are small, suggesting that AS selects attribute sets sufficiently enough to maintain the accuracy for Ant-Miner and C4.5.

Keywords: Attribute Selection, Ant Miner, Air pollution, Hospital admission.

1 Introduction

Attribute Selection (AS) methods are generally applied to practically reduce computation time on large data sets, e.g., the thousands of attributes that are often seen in text or web mining problems, whereas attribute sets in environmental science can be reasonably small, e.g., from ten to less than hundreds, due to availability and accessibility, e.g., historically unmeasured variables or limitations in setting up monitoring sites. AS acts as a knowledge discovery tool by ranking the importance of attributes (ranking filter method) or selecting a subset of attributes (subset evaluator) [1, 2, 3]. This study uses four commonly known AS methods in WEKA [4]; Information Gain Attribute Ranking (IG) [5], Relief (RLF), e.g., [6], Correlation-based Feature Selection (CFS) [2], Consistency-based Subset Evaluation (CNS) [7] and a newer AS method; Tree Node Selection (TNS) for unpruned (TNSU) and for pruned (TNSP) [1], to identify fewer but key climate and air pollution factors to understand the underlying structures of acute respiratory hospital admission patterns.

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 853–860, 2009.

Performance of each selected attribute set (input) was then tested with the Ant-Miner [8] and C4.5 [5] algorithms to observe prediction ability of admission pattern (class; *yes* for two or more admission counts otherwise *no*, which follows *outbreak* definition [9]). While the acute respiratory admissions are commonly known to be affected by changes in various air pollution and climate levels and generally investigated by statistical analysis [10], the goal of this study is to determine key climate and air pollution factors for respiratory hospital admissions from applying AS methods rather than reducing the attribute dimensions, which is a new approach in studying air pollution and health. To the best of our knowledge, these data mining attribute selection methodologies have not been commonly examined as data pre-processing methodologies or for selection of key attributes to obtain improved results for the Ant-Miner classifier. In particular, solving environmental science problems by using a combination of Ant-Miner and attribute selection is not yet a common procedure.

The next method section briefly outlines each AS method, followed by previous AS in Ant-Miner, the study data, in addition to introducing the motivation for producing the admission prediction model for the study site, and how AS methods would help. The final two sections present results and discussions, and summarize the finding as conclusions. The aim of this paper is to introduce applications of various AS methods with Ant-Miner and C4.5 algorithms for a real environmental science case study. The brief descriptions of AS, Ant-Miner and C4.5 algorithms are discussed or omitted as details are shown in [1, 2, 4, 5, 8, 11].

2 Methods

2.1 Background of Attribute Selection Methods

Detailed performance of AS methods, e.g., classification accuracy and reduction of attributes, that were used in this study were previously investigated by [1] for testing IG, RLF, CFS, CNS and wrapper methods to propose TNSP and TNSU, and [2] for testing IG, RLF, CNS, principal component analysis and wrapper methods to propose CFS, from applying naïve Bayes and C4.5 algorithms on various sized benchmarking databases. For example, reference [1] tested 33 benchmark datasets from 4 attributes to less than 300 attributes and from 40 to less than 50,000 instances. The selected attribute set varies depends on how each algorithm works; see details in [1, 2, 4]. The simplest attribute selection method [2] is IG [5], which quickly selects and orders attributes by importance by measuring the information gain in respect to the class, RLF [6] searches important attributes by repeatedly selecting a randomly selected instance from its two nearest neighbours between the same class and others, CNS [7] looks for a subset of attributes with the best consistency in class values, CFS [2] searches for a good subset of attributes by considering the usefulness of individual features at predicting each class, and a newer attribute selection method, Tree Node Selection (TNS), identifies a set of attributes using a pre-constructed decision tree as an information source by counting numbers of instances that go through nodes and classes. For example, the root node of a decision tree tends to be most important as it connects to the rest of the nodes to distribute instances to classes, but if the attribute was more frequently used to construct many leaf nodes, then that attribute can also be important; see details in [1].

From the benchmark experiments, references [1] and [2] found that the wrapper is the best AS method, but it is time consuming. Reference [1] found that TNS performed consistently in reduction of attributes and obtained high accuracy over various data mining attribute selection approaches, whereas other methods tend to trade off performance in reduction of attributes and accuracy. Reference [2] found that CFS, CNS and RLS are good overall, but there is a trade off in performance among them.

2.2 Attribute Selection with Ant-Miner

The Ant Colony Optimization (ACO) algorithm is a swarm intelligence technique that mimics real ant behaviour. Recently, ACO has been used to solve attribute or feature selection problems. Reference [12] used ACO in rough set theory [13] to obtain high accuracy and minimum sets of features, since ACO finds solutions rapidly with very small cardinality during its pheromone update rule and solution construction process. Reference [14] developed an ACO feature selection method and tested it on a text categorization problem against other data mining and statistical attribute selection methodologies, Information Gain (IG), χ^2 statistics (CHI) and genetic algorithms (GA), by performing nearest neighbour classification. They found that ACO outperformed IG and CHI, and GA is almost comparable to ACO in terms of maintaining the accuracy and selecting minimum feature subsets. They stated that for datasets with more features, ACO has a strong search capability in the problem space, as a search continues until the optimal solution is found, whereas GA cannot find a better one after finding a sub-optimal solution.

2.3 Ant-Miner and C4.5 Classifiers

In order to test performance of attribute sets selected by AS methods, Ant-Miner and C4.5 algorithms were used to compare the prediction ability of admission counts using the smaller sets of selected attributes. Since this study is a preliminary investigation, the traditional Ant-Miner [8] has been used, because it is still a flexible and robust classification mining method which works well [15, 16], even though newer Ant-Miner algorithms have been developed, e.g., Ant-Miner 2 [17], ACO-Miner [11] and TACO-Miner [16]. In comparison to Ant-Miner, one of the most well known classification algorithms, C4.5 [5], was tested because Ant-Miner [8] is similar to a decision tree algorithm that discovers classification rules by following a divide-and-conquer approach:

IF < term1 and term2 and ...> THEN <class>

However, the heuristic functions for decision tree algorithms and Ant-Miner differ in how they consider the entropy; for the former they are computed for an attribute as a whole, but the latter computes them for an attribute-value pair only [8]. For Ant-Miner and C4.5 used the default parameter setting of Ant-Miner software [8] and WEKA [4] respectively was used.

2.4 Attribute Selection Process

Fig. 1 describes attribute selection steps. Firstly, the entire data (full attribute set) was divided into 90% (from the start of the studied period) and 10% (towards the end of studied period) to create the training and test set. Ant-Miner and C4.5 classifiers are applied on the training set to obtain the *original* classification accuracy (before the AS process) via the 10-fold cross validation process. Secondly, two sets of AS approaches are carried out. Ranking filter approaches (TNSP, TNSU, IG and RLF) ranks each attribute by its importance, i.e., the top labelled rank from "1" indicates the most important attribute and so on. Each set of ranked attributes separately runs the Ant-Miner and C4.5 classifiers, iteratively removing the least important attribute one by one to obtain the classification accuracy and the process continues until a single attribute remains. Subset evaluator approaches (CFS and CNS) select the attribute set at once, whereas the attribute set that obtained the highest accuracy for the ranking filter is used for the prediction process. Ant-Miner and C4.5 are separately run with the selected subset of attributes from the training set to extract classification rules. The created rules are then tested on the test set (unknown data points, not used to select the attribute set) to obtain the prediction accuracy, the *final* classification accuracy, of the respiratory admission pattern.

To identify key climate and air pollution factors for the admission pattern, the obtained attribute set (with the highest classification accuracy) is examined and compared among AS methods. Here, the top 3 commonly selected attributes throughout all AS methods will be summarised as follows. When TNSU ranks TG at rank 1, it gives one point to TG. If another AS method ranked TG at rank 2, it adds another point to TG. The total points are added up. The attribute that records the highest score is considered to be the most frequently selected attribute over all AS methods. Note that all points are counted equally as "one point" regardless of rank, i.e., first or third rank.

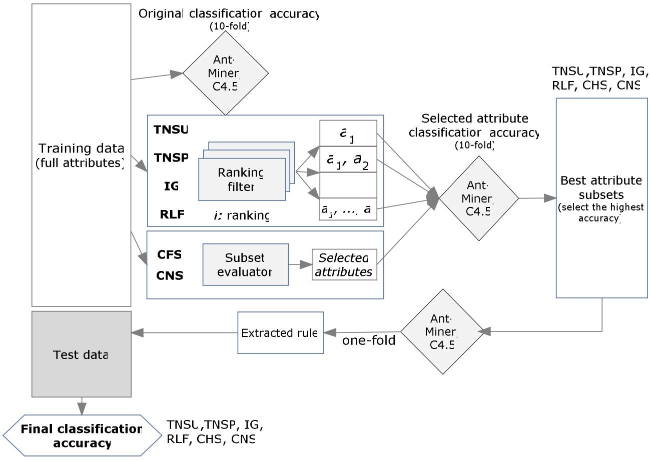


Fig. 1. Attribute selection processes for Ant-Miner and C4.5 classifiers

Relative proportion of selected number of attributes (in %) is calculated to compare the attribute reduction performance among all AS methods. The two sample means (and standard deviation) of the *original* and *final* classification accuracy of Ant-Miner and C4.5 are calculated to assess how the means of classification accuracy differ among classification algorithms.

2.5 Studied Data

The study area, Christchurch, New Zealand, suffers from severe air pollution problems in winter due to domestic heating by burning wood, e.g., [10]. The studied area, Christchurch City, is located in the South Island of New Zealand. The main winter air pollutants in Christchurch are carbon dioxide (CO_2) from domestic heating and motor vehicles, particulate matter (PM and PM_{10}, particles of diameter 10 micrometers or less) from domestic heating, sulfur dioxide (SO_2) from industry, e.g., [10]. Some pollutants can record beyond the acceptable air pollution guideline during winter due to the heavy use of wood for domestic heating. It is desirable to promote a good prediction method for the outbreak acute admission rate in order to help with the hospital care management.

A total of 16 daily measurements was collected over a four year period (October 1998-September 2002) from a single air pollution monitoring site, located in a medium-size residential area (see details in [10]) in northern Christchurch City, of air pollution and climate is investigated; PM_{10}, SO_2, CO_2, relative humidity (RH), an indication of the temperature inversion formation (calculated from the difference between the temperatures at 1m and 10m above the ground, with negative values indicating temperature inversion formation, labelled separately as TG, TT and TD), wind speed (WS), wind direction (Wdir), atmospheric pressure (P), radiation hours (Rad), sunshine hours (Sun), rainfall (Rain), maximum and minimum daily temperature and the average of these (TMax, TMin and TAv). All air pollution and climate data were scaled (no specific units). Over the same period, daily counts of acute hospital admissions due to respiratory system problems (ICD-9: 460-519) were obtained for residents domiciled within 2 km (age 0-98 years, n=878 for female, and n=1061 for male) of the air pollution monitoring site. The studied data contained a maximum of about 4% missing values, mainly from $\mathbf{SO_2}$ and temperature inversion data points, but were separately imputed and did not significantly alter results.

3 Results and Discussion

3.1 Key Attributes for the Admission Outbreak

Table 1 shows a summary of numbers and relative reduction (in %, with higher proportions indicating greater reduction) of selected numbers of attributes. TNSU selected only three attributes (minimum subset and 81.3% reduction) whereas CNS selected 11 attributes (maximum subset and 31.3% reduction). TNSU and TNSP selected the smallest numbers of attributes. The rest of the AS methods selected more than 7 out of 16, so about half of the attributes were removed. Table 1 also shows a summary of selected attributes and an assessment of top 3 selected attributes over all AS methods. The TNSU selected TG, CO and RH. The top 3 most commonly selected attributes over all AS methods are CO and RH, which are ranked highest (4 times) and followed by SO_2 and TG (3 times). It could be said that these four attributes are key factors that can help improving or are underline potential factors of the admission prediction. Additionally, all three attributes selected by TNSU are three of the four most commonly selected attributes by all other AS methods.

3.2 Selection of Attributes for Ant-Miner and C4.5

In Table 2, the means of Ant-Miner and C4.5 from the training sets shows the mean of Ant-Miner (μ=65.8) is not significantly larger, or even equal (p=0.05 for one-tailed using assuming equal variances, as F-test for equal variances shows p=0.13) to C4.5 (μ=65.0). While using all attribute sets recorded similar classification accuracies (Ant-Miner; 64.5% and C4.5; 63.8%), the removal of attribute was not carried out to significantly improve the classification accuracy. However, the proportion of attribute removal was significantly successful (up to 81% for TNSU) and the quality of classification accuracy was maintained, even with much smaller attribute sets.

Although, overall Ant-Miner classification accuracy (training set) recorded slightly higher classification accuracy with the ranking filter approaches; RLF (66.3%), TNSP

Table 1. Numbers of selected attributes, relative attribute reduction (in %) and a summary of selected attributes

Proportion of original class: yes 49% no 51%	TNSU	TNSP	RLF	IG	CFS	CNS	Full	Attribute frequency (Top 3)
# of selected attributes	**3**	5	7	9	8	11	16	
Reduction of attributes (%)	**81.3**	68.8	56.3	43.8	50.0	31.3	0.0	
Ranking of attributes								
1 (most important)	TG	TG	RH	TG	SO_2	SO_2	SO_2	**3**
2	CO	RH	TD	TT	CO	CO	CO	**4**
3	RH	CO	SO_2	Rad	RH	PM_{10}	PM_{10}	1
4		SO_2	TD	Tmin	TG	RH	RH	**4**
5		Tmax	Tav	TD	TD	TG	TG	**3**
6			Rad	Tmax	Rad	TT	TT	1
7			Tmax	Tav	Tmax	TD	TD	1
8				SO_2	Tmin	Rad	WS	
9				RH		Tmax	Wdir	
10						Tmin	P	
11						Tav	Rad	1
12							Sun	
13							Rain	
14							Tmax	
15							Tmin	
16 (the lowest ranking)							Tav	

Table 2. Summary of classification accuracy for training and test sets for Ant-Miner and C4.5 classifiers

Classification	Full (before AS)	TNSU	TNSP	RLF	IG	CFS	CNS	Mean	SD	Two-sample means of Ant-Miner and C4.5
Ant-Miner (training)	64.5	66.0	66.2	**66.3**	66.1	64.8	65.5	65.8	0.6	p =0.05 (one-tail)
C4.5 pruned (training)	63.8	65.8	65.9	65.9	64.0	64.4	63.8	65.0	1.0	
Ant-Miner (test)	54.3	**59.3**	55.7	55.0	47.9	54.3	56.4	54.8	3.8	p =0.37 (one-tail)
C4.5 pruned (test)	54.3	57.9	54.3	55.0	56.4	54.3	54.3	55.4	1.5	
Ant-Miner rules	7	7	7	7	7	8	8			
C4.5 leaves	7	4	7	3	4	8	7			
C4.5 size of tree	13	7	13	5	7	15	13			

(66.2%), IG (66.1%) and TNSU (66.0%), compared with the subset evaluator; CNS (65.5%) and CFS (64.8%). Similarly, overall C4.5 classification accuracy (training set) recorded similar classification accuracies, but also filter AS approaches; RLF and TNSP (65.9%), and TNSU (65.8%) except IG (64.0%) performed slightly better than subset evaluator approaches CFS (64.4%) and CNS (63.8%). Note that Table 2 also shows the information about the constructed rule, e.g., size of tree or number of rules, but are not specifically discussed.

3.3 Testing Prediction Performance with Selected Attributes for Ant-Miner and C4.5

Results of predicted admission accuracy are also summarised in Table 2. The two means of Ant-Miner and C4.5 using the test sets suggests that there is no significant evidence to say that the mean of C4.5 prediction accuracy (μ=55.4%) over all AS methods are higher than or equal to Ant-Miner (μ=54.8%) (*p*=0.37 for one-tail using assuming unequal variances, as F-test for equal variances shows *p*=0.03). While the original class proportion of *yes* is 49% and *no* is 51%, it could be said that both classifiers are slightly more effective than just guessing either class, especially TNSU for Ant-Miner, which achieved the highest prediction accuracy around 60%. In fact, TNSU also obtained the highest prediction accuracy for C4.5 (57.9%). On the other hand, IG recorded the lowest prediction accuracy for Ant-Miner (47.9%) and CFS and CNS recorded the lowest for C4.5 (54.3%). As previously mentioned, the AS method may be slightly more effective on Ant-Miner than C4.5.

Even though the classification accuracy was not significantly improved with fewer attributes, a possible reason why TNSU provides the highest prediction accuracy for Ant-Miner over C4.5 can be considered that TNS searches important attributes by assessing the connectivity of adjacent nodes in the decision trees; frequently connected pairs of attributes in the decision tree are more important than ones that are not connected. Ant-Miner rules are constructed by pheromone trails, which produce high solutions based on a high probability pair of attributes during updating pheromone iteratively as ants write, read and estimate the amount of pheromone trail to build a good solution [11]. Hence, the attributes that are selected by TNS may strengthen the search between attributes for Ant-Miner because Ant-Miner similarly searches attributes that have higher probability between nodes (pheromone trial). Providing fewer but specifically selected representative attributes may help increase efficiency in finding a solution in the Ant-Miner in less confusing manners. Whereas IG measures the information gained with respect to class, it may not directly consider the strengths between attribute nodes. Surprisingly, CFS produced the same classification accuracy regardless of removing or full attributes sets for both Ant-Miner (54.3%) and C4.5 (54.5%), even though CFS selects individual features at predicting each class along with the level of inter-correlation [2], which could strength the path that was taken during the Ant-Miner search. The studied data set may not have such good level of inter-correlation.

4 Conclusions

This paper examined six different data mining attribute selection (AS) methods, TNSU, TNSP, RLF, IG, CNS and CFS, to extract key climate and air pollution factors by predicting the acute respiratory admission counts with Ant-Miner and C4.5

algorithms. TNSU preformed best to remove up to 80% of the attributes by selecting only three attributes; temperature at ground level, carbon monoxide and relative humidity, and obtained a classification accuracy improvement (from 2% to 5%) for both Ant-Miner and C4.5. All other AS methods removed approximately half of the attributes, seem to trade off between attribute reduction performance and maintaining prediction accuracy. This is a preliminary experiment using data mining AS methods on environmental and health study with Ant-Miner. It will be expected to keep investigating other newer Ant-Miner algorithms, such as Ant-Miner 2, ACO-Miner, and TACO-Miner, with much larger attribute sets in future.

Acknowledgments. Thanks to Mr. P. Pearson for editorial work and the reviewers for useful comments. Data were provided by ECan, University of Canterbury (Dept. of Physics) and NIWA, and CDHB (URB/06/05/031).

References

1. Fukuda, K., Martin, B.: Decision Trees as Information Source for Attribute Selection. In: SSCI 2009 IEEE CIDM, pp. 101–108 (2009)
2. Hall, M.A., Holmes, G.: Benchmarking Attribute Selection Techniques for Discrete Class Data Mining. IEEE Trans. Knowl. Eng. 15, 1437–1447 (2003)
3. Jensen, R., Shen, Q.: Fuzzy-Rough Sets Assisted Attribute Selection. IEEE Trans. Fuzzy Syst. 15, 73–89 (2007)
4. Witten, I.H., Frank, E.: Data mining: Practical machine learning tools and techniques with Java implementations, 2nd edn. Morgan Kaufmann, San Francisco (2005)
5. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo (1993)
6. Kira, K., Rendell, L.: A practical approach to feature selection. In: 9th ICML 1992, pp. 249–256 (1992)
7. Liu, H., Setiono, R.: A probabilistic approach to feature selection: a filter solution. In: Proc. 13th ICML 1996, pp. 319–327 (1996)
8. Parpinelli, R.S., Lopes, H.S., Freitas, A.A.: Data Mining With and Ant Colony Optimization Algorithm. IEEE Transactions on Evolutionary Computing 6, 321–332 (2002)
9. PEHG, Annual summary of outbreaks in New Zealand, Population and Environmental Health Group Institute of Environmental Science and Research Ltd. (2007)
10. Fukuda, K., Takaoka, T.: Analysis of Air Pollution (PM10) and Respiratory Morbidity Rate using K-Maximum Sub-array (2-D) Algorithm. In: SAC 2007, pp.153–157 (2007)
11. Dorigo, M., Stützle, T.: Ant Colony Optimization. MIT Press/Bradford Books, Cambridge (2004)
12. Ke, L., Feng, Z., Ren, Z.: An efficient ant colony optimization approach to attribute reduction in rough set theory. Pat. Rec. Let. 29, 1351–1357 (2008)
13. Pawlak, Z.: Rough sets: theoretical aspects of reasoning about data. Kluwer, Boston (1991)
14. Aghdam, M.H., Ghasem-Aghaee, N., Basiri, M.E.: Test feature selection using ant colony optimization. Expert Sys. Appl. 36, 6843–6853 (2008)
15. Wang, Z., Feng, B.: Classification rule mining with an improved ant colony algorithm. In: Webb, G.I., Yu, X. (eds.) AI 2004. LNCS (LNAI), vol. 3339, pp. 357–367. Springer, Heidelberg (2004)
16. Thangavel, K., Jaganathan, P.: Rule Mining algorithm with a new ant colony optimization algorithm. In: Int'l conf on computational intelligence and multimedia applications, pp. 135–140 (2007)
17. Liu, B., Abbas, H.A., Mckay, B.: Classification rule discovery with ant colony optimization. In: IEEE/WIC Intelligent Agent Technology, pp. 83–88 (2003)

Combined Unsupervised-Supervised Classification Method

Urszula Markowska-Kaczmar and Tomasz Switek

Wroclaw University of Technology, Wroclaw, Poland
urszula.markowska-kaczmar@pwr.wroc.pl
http://www.iis.pwr.wroc.pl/~kaczmar/

Abstract. In the paper a novel method of classification is presented. It is a combination of unsupervised and supervised techniques. First, the method divides the set of learning patterns into smaller ones in the clustering process. At the end of this phase a hierarchical structure of Self Organizing Map is obtained. Then for the leaves the classification rules are searched. To this end Bee Algorithm is used. The accuracy of the method was evaluated in an experimental way with the use of benchmark data sets and compared with the result of other methods.

1 Introduction

Classification techniques are key elements in solving problems in a number of disciplines. That is why with the rapid development of computer technology, many classification methods have been developed and used: SVMs [7], decision trees [6], neural networks [3], k-Nearest Neighbours (k-NN), Naive Bayes (NB) [6], rule based classification methods (for instance FOIL)[6] and others. They differ in the accuracy and ability of explanation of the classification decision. For numerical attributes some of them are strongly dependent on the discretisation method applied in preprocessing step. Generally, SVMs and neural networks tend to perform much better when dealing with multidimensions and continuous features. For these models a relatively large sample size is required in order to achieve its maximum prediction accuracy whereas NB may need a relatively small data set. Naive Bayes and the k-NN can be easily used as incremental learners whereas rule algorithms cannot. Some algorithms, for instances Naive Bayes is naturally robust to missing values since these are simply ignored in computing probabilities. On the contrary, k-NN and neural networks require complete patterns (without missing values in vectors). Moreover, k-NN is generally considered intolerant of noise because its similarity measure strongly depends on errors in attribute values, thus leading it to misclassify a new instance on the basis of the wrong nearest neighbours. Contrary to k-NN, rule based methods and most decision trees are considered resistant to noise because their pruning strategies avoid overfitting the data in general and noisy data in particular. More discussion about virtues and shortcomings of the methods can be found in [5].

In this paper, we have proposed a novel classification method which is a combination of unsupervised and supervised approaches. Its origin comes from

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 861–868, 2009.

obvious assumption that to solve smaller problem is much easier. That is why in the first phase the clustering is performed giving as the result subsets of patterns for which supervised classification is performed. The next section gives the general view of the method. Its details are presented in the section 3. Experimental study which aim was to evaluate the method is shown in section 4. The paper ends with conclusion summarizing the paper and describing the future plans.

2 CUSC Method – General Overview

CUSC is a novel inductive method of building a classifier. It uses both supervised and unsupervised learning paradigms. The main feature of CUSC is that it divides set of learning patterns into smaller ones in the clustering process using approach similar to Self Organizing Map but the structure of the network has a hierarchical form (a tree). This makes the supervised part of building CUSC easier, as the process is handling less complex sets at the time step.

A creation of the classifier has two phases. In the first phase a clustering method is applied. It is performed with the use of hierarchical neural network which adapts itself to the presented patterns. During training its structure is hierarchical and dynamic assuming form a tree (Fig.1) which is composed of both – structural elements that do not directly influence on classification decisions and leaves that a role is to classify patterns. These classifying elements of CUSC can be divided into two distinct types called *simple form* and *complex form*. The elements type is defined as a *simple form* when after building the structure the training patterns, assigned to it, are belonging to one class only.

The *complex form* elements after clustering contain the set of patterns from more than one class. In other words, the simple form classifying elements contain patterns with the label of a single class. For the *complex form* elements

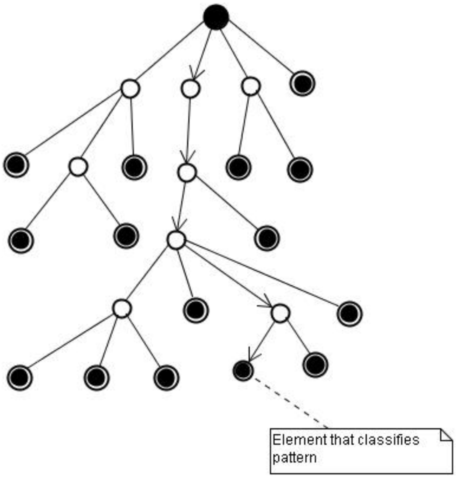


Fig. 1. The tree structure used in CUSC

classification proceeds according to the set of conjunctive rules that are searched in the second phase of the method. This set is used for determining appropriate label for the pattern that is to be classified after building classifier. The method of rules extraction for *complex form* elements in the structure is described in the subsection 3.2.

When the classifier is trained the process of decision making in CUSC consists of two stages. The first one needs to find appropriate classifying element in the tree and in the second one the classification decision is made by the classifying element chosen in the previous stage. The whole classification process is performed in four steps:

1) Initialization: set E_p as the root node of a tree and as **x** the pattern to be classified.
2) If E_p has any descendent nodes go to step 3) otherwise go to step 4).
3) Select a descendent node that presents the highest resemblance to the pattern **x** according to the similarity measure (for example the smallest Euclidean distance). Set as the E_p the selected descendent node and go to step 2).
4) Classify the pattern **x** and return a label of a class as a result.

The last step, which performs the classification process is dependent on the type of classifying element that it is applied to. In case of *simple form* classifying elements classification is made by returning the label that the element holds. When the classifying element is of *complex form*, the classification is performed by making a decision based on the set of rules in the form of the IF...THEN (eq. 1).

$$IF\ prem_1\ AND\ prem_2\ AND\ ...\ AND\ prem_m\ THEN\ c_i \tag{1}$$

where premise $prem_i$ expresses a condition imposed on the value of attribute x_i. This condition must be satisfied by the pattern to classify it to class c_i.

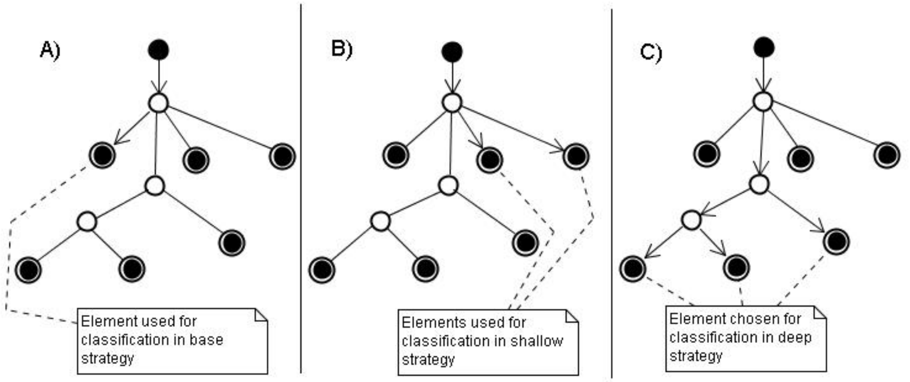

A)
Element used for classification in base strategy
B)
Elements used for classification in shallow strategy
C)
Element chosen for classification in deep strategy

Fig. 2. Strategies of surveying the : A) basic strategy; B) *shallow* strategy (when A) fails); C) *deep* strategy (when A) and B) fail)

In the case that the set of rules cannot be applied to the pattern (no rule is fired for it), the CUSC is trying to find the solution in the neighbourhood of the chosen classifying element. The method applies two strategies of surveying the neighbourhood: *shallow* strategy and *deep* strategy (Fig. 2). The shallow strategy consists in trying to classify the pattern by the classifying elements that are also descendent nodes of the parent of classifying element chosen in the sequence. If the shallow strategy fails in giving the result (lack of classification), the deep strategy is used. The deep strategy consists in determining the most common answer that is given by structural elements of the parent of classifying element chosen in the sequence. The answers given by structural elements are cumulated (the most common answer is chosen) from the sub-trees of these elements. In case that both strategies fail, the CUSC returns an empty label meaning that the classifier doesn't know the answer.

3 The Method in Detail

The process of creating our hybrid classifier can be divided into two phases: building a classifier structure, and a searching for a set of classification rules for complex form elements.

3.1 The First Phase – Building the Structure

This process is unsupervised and proceeds accordingly to the following sequence of steps:

1. Initialization: set the root of a tree as E_p and the initial learning set as X_p .
2. Check whether E_p is fulfilling the STOP condition. If it is TRUE then go to second phase, otherwise go to step 3.
3. Perform clustering of the X_p set. Set the acquired clusters as the children nodes of E_p and assign to them the sub-sets of X_p that they cover. If the clustering results give one cluster go to second phase.
4. For each of the children nodes of the E_p go to step 2. with treating the child node as E_p, and assign to X_p the set of training patterns that belong to E_p.

The STOP condition from the step 2. of this algorithm is preventing the overgrowth of the tree structure. To fulfil the STOP condition, the number of patterns in the learning sub-set assigned to E_p must be smaller than the preset threshold, which dependent on the number of patterns in the initial learning set.

The clustering in CUSC is performed with the use of *Neural Clustering* (NC) method. NC is strongly related to SOM network. However, in contrary to SOM networks, NC is dynamically building the clusters structure allowing for their fast growth, and then gradual degradation. In this method the temperature and the mass of neurons are introduced. They have an influence on similarity and weight modification functions. With the timestep k the temperature decreases according to the eq.2.

$$T(k+1) = T(k) - \frac{1}{2} \cdot \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad (2)$$

where

$$x = \frac{1}{E(k/2)+1}; E - Entier function.$$

The role of the temperature is to affect an influence between neurons, which is realized by the weight modification (eq. 6). It enables NC method to create expanded sets of clusters in the early phase of clustering while the temperature is high, and allows gradual degradation providing stabilization of clusters in the final steps of the method.

Each of the structure neurons is characterized by its mass M_n. It assigns the relative number of patterns clustered by a neuron (a ratio of Z_n to Z – which is the number of all patterns used in the clustering process), eq. 3.

$$M_n = \frac{card(Z_n)}{card(Z)} \tag{3}$$

The neurons mass determines its behaviour in the learning process. The greater the mass is, the stronger the neuron affects others and it is less affected by them. Because the structure of clusters is dynamic and the patterns set assigned to each neuron can change over the time, there is a possibility that neuron mass will reach 0. In this case the neuron is removed from the structure. This results in degeneration of the tree. During training process, after each pattern presentation the distance D between weights $\mathbf{w_n}$ of n-th neuron and the given pattern $\mathbf{x}$ is calculated according to the eq.4.

$$D\big(\mathbf{w_n}(t), \mathbf{x}(t)\big) = F\big(\mathbf{w_n}(t), \mathbf{x}(\mathbf{t})\big) - M_n(k)F\big(\mathbf{w_n}(t), \mathbf{x}(\mathbf{t})\big) \tag{4}$$

where F is an Euclidian distance between the weights $\mathbf{w_n}$ and the given pattern $\mathbf{x}$, M_n is a mass of n-th neuron. In case the greatest distance is greater than an average distance between weights of existing neurons a new neuron is added to the structure. Its weights are set as equal to the given pattern. Otherwise the weights of existing neurons are updated. For the winning neuron (the neuron for which the distance D is the smallest one) the weights $\mathbf{w}_w$ are changed as follows:

$$\mathbf{w_w}(t+1) = \mathbf{w_w}(t) + T(k)\big(\mathbf{x}(t) - \mathbf{w_w}(t)\big)\big(1 - M_w\big) \tag{5}$$

The weights of neurons in the neighbourhood are changed according to the eq. 6.

$$\mathbf{w_n}(t+1) = \mathbf{w_n}(t) + \frac{T(k)\big(\mathbf{x}(t) - \mathbf{w_w}(t)\big)\big(1 - M_w\big)}{D\big(\mathbf{w_n}(t), \mathbf{w_w}(t+1)\big)} \tag{6}$$

In the next step the mass of the winning neuron is updated. In each epoch during training the temperature is updated, as well. The end of the first phase results in the clustered training patterns that create the patterns sets mutually exclusive.

3.2 The Second Phase – Searching for Classification Rules in the Complex Form Elements

As it was mentioned, the complex form elements cluster patterns that belong to more than one class. For such cases a set of conjunctive rules in the attributes

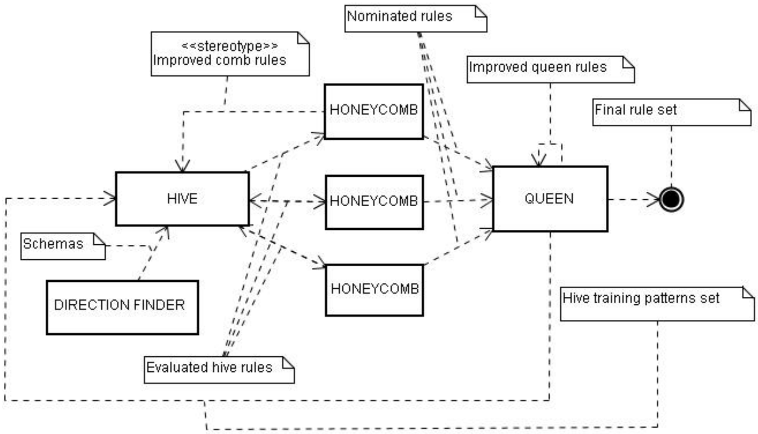


Fig. 3. Bee Miner structure and flow of rules between key elements

logic is searched by a rule extractor (Bee Miner). It is composed of four key elements (Fig. 3): *direction selector*, *hive*, *honeycombs*, and *queen*.

The *direction finder* is used only at the moment of creating new rules. Its task is to pick an appropriate set of attributes (schema) on which the new rules will be created. It is realized on the basis of evolutionary algorithm, where rank selection, mutation and homogenous crossover are applied. A chromosome consists of the list of attributes. The role of this algorithm is to search for subspaces where patterns from different classes can be easily separated. Individuals are evaluated on the basis of the density metrics. To calculate this value the gravity center for each class must be found and the average distance between patterns of this class and the center is computed next. Then these values are compared for pairs of classes in order to determine whether the classes are separated. The fitness value of an individual is based on the relative number of separated pairs of classes.

The purpose of a *hive* is creation and a rough evaluation to which honeycomb the premise part of a rule fits the best. The premise part is based on the schema selected in the direction finder mechanism. As the result of this step a list of input features is obtained in the form of a vector: $\mathbf{x}$=[x_1,x_2,..., x_k] where $k <= n$ (n is the dimension size for the classification problem). Then the schema is converted to the premise part of a rule by applying random operator and the value from the attribute domains. Generally, we can say the i premise has a form x_i R $value_k$, where x_i is a variable representing i-th attribute, R stands for an operator ($R \in \{<, >, =\}$) and $value_k$ is one of the possible values for this attribute. After the process of a rule evaluation the hive transfers newly created rules to appropriate honeycombs.

Each *honeycomb* represents one class, for which it tries to find the best set of rules. The mechanism in honeycomb is realized by a limited queue that gathers and orders the best rules found by hive. Each rule in the queue has a nomination counter that informs how many times the rule was in the queue in the honeycomb. If the counter reaches the appropriate value, the rule migrates to the set of nominated rules. These rules are collected by a queen.

The queen is re-evaluating and selecting rules in order to create the set of rules with the best accuracy. The rules taken to re-evaluation and selection are from the gathered nominated rules sets and from earlier created set of rules. The selection process of a new temporal solution relies on the choice of rules that are characterized by the highest accuracy.

4 Experimental Study

The classifier evaluation is based on the prediction accuracy (the percentage of correct predictions divided by the total number of predictions). The experiments were performed for four benchmark data sets taken from [1]. The results are compared with SVM and C4.5 results published at [2].

The characteristics of the applied data sets are presented at the bottom part of Table 1. As it can be noticed they contain a various number of attributes and a number of classes. At the top part of this table the results are shown. They represent an average from ten runs performed with the use k cross validation method with k=10. Throughout testing of the CUSC method two sets of parameters were used. The first one supported rules extraction (LGT set) by limiting the growth of CUSC structure. It increased the number of patterns for each single extraction. The second set of parameters (ETS set) limited rules extraction by building expanded tree structure. The research have shown that the quality of CUSC method results is strongly dependent on the parameter set. The differences between the results with the use GTS and ETS sets were reaching up to 8%. The best presented example are the results acquired for classification of Iris set. In that case the usage of ETS leads to creating structures without any classification elements in complex form. It effected in giving 97 % accuracy, while when the ETS set of parameters were used the accuracy reached only 90 %. On the other hand the situation of performing classification for the Wine data was opposite. In this case GTS set of parameters has given better results.

Table 1. The classification accuracy for CUSC and other methods for benchmark data sets; (WBC) is the abbreviation from Wisconsin Breast Cancer

Comparison with other methods in terms of accuracy				
	(WBC)	Sonar	Wine	Iris
CUSC	0.95 (ETS)	0.82(ETS)	0.94(LGT)	0.97(ETS)
SVM	0.97	0.82	0.97	0.96
C4.5	0.95	0.76	lack of data	0.95
data sets characteristics				
Pattern number	683	208	178	150
Class number	2	2	3	3
Attribute number	9	60	13	4

5 Conclusion and Future Work

For the tested data sets the results included in Table 1 show that CUSC has accuracy similar to SVM, and better than C4.5 classifier. But the key question when dealing with machine learning classification is not whether a learning algorithm is superior to others, but under which conditions a particular method can significantly outperform others on a given application problem. For this reason there is a need for more experimental study with the method. Nonetheless the results acknowledge relevance of the further works on CUSC.

However the current version of our classifier does not give uniform representation of all decision nodes the original goal was to obtain classification rules for all leaves. This assumption will be realized in the future. The further development of the method will be focused on decreasing a complexity of the method and on adjusting mechanisms of rules improvement in the Bee Miner method. Additionally, mechanism of local improvement of sub-trees is under development. Its purpose is to change or re-assemble those parts of tree structure that have negative influence on the classification accuracy.

References

1. Asuncion, A., Newman, D.J.: UCI Machine Learning Repository, University of California, Irvine, School of Information and Computer Sciences (2007), http://www.ics.uci.edu/~mlearn/MLRepository.html
2. Duch, W.: Datasets Used for Classification: Comparison of Results (2009), http://www.is.umk.pl/projects/datasets.html
3. Haykin, S.: Self-organizing maps. Neural networks - a Comprehensive Foundation, 2nd edn. Prentice-Hall, Englewood Cliffs (1999)
4. Karaboga, D.: An Idea Based On Honey Bee Swarm for Numerical Optimization. Technical Report-TR06, Erciyes University, Engineering Faculty, Computer Engineering Department (2005)
5. Kotsiantis, S.B.: Supervised Machine Learning: A Review of Classification Techniques. Informatica 31, 249–268 (2007)
6. Mitchell, T.: Machine Learning. McGraw-Hill, New York (1997)
7. Sullivan, K., Luke, S.: Evolving Kernels for Support Vector Machine Classification. In: Genetic And Evolutionary Computation Conference Archive Proceedings of the 9th Annual Conference on Genetic and Evolutionary Computation, pp. 1702–1707 (2007)

Author Index