

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

LNAI 5711

Knowledge-Based and Intelligent Information and Engineering Systems

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

1
Part I



 Springer

Lecture Notes in Artificial Intelligence 5711

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 I

Series Editors

Randy Goebel, University of Alberta, Edmonton, Canada
Jörg Siekmann, University of Saarland, Saarbrücken, Germany
Wolfgang Wahlster, DFKI and University of Saarland, Saarbrücken, Germany

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 3-642-04594-4 Springer Berlin Heidelberg New York
ISBN-13 978-3-642-04594-3 Springer Berlin Heidelberg New York

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable to prosecution under the German Copyright Law.

springer.com

© Springer-Verlag Berlin Heidelberg 2009
Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 12763962 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

Organization

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. Ríos
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, Katsuhiko	Osaka Prefecture University, Japan
Hong, Tzung-Pei	National University of Kaohsiung, Taiwan
Hornig, 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
Pratihari, 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
Ushiana, 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	Miroslav Karny	Tuan Pham
Bojana Basic	Honghai Liu	Bernd Reusch
Floriana Esposito	Ngoc Nguyen	Sebastián Ríos
Anne Håkansson	Andreas Nuernberger	

Invited Sessions Programme Committee

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

KES 2009 Reviewers

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

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

Fuzzy and Neuro-Fuzzy Systems

Intersection Search for a Fuzzy Petri Net-Based Knowledge Representation Scheme	1
<i>Slobodan Ribarić, Nikola Pavešić, and Valentina Zadrija</i>	
Parametric Uncertainty of Linear Discrete-Time Systems Described by Fuzzy Numbers	11
<i>Petr Hušek</i>	
A Flexible Neuro-Fuzzy Autoregressive Technique for Non-linear Time Series Forecasting	22
<i>Alejandro Veloz, Héctor Allende-Cid, Héctor Allende, Claudio Moraga, and Rodrigo Salas</i>	

Agent Systems

Multiagent Security Evaluation Framework for Service Oriented Architecture Systems	30
<i>Grzegorz Kołaczek</i>	
Describing Evolutions of Multi-Agent Systems	38
<i>Sergey Babenyshev and Vladimir Rybakov</i>	
Functionality and Performance Issues in an Agent-Based Software Deployment Framework	46
<i>Mario Kusek, Kresimir Jurasovic, and Ignac Lovrek</i>	
A Consensus-Based Integration Method for Security Rules	54
<i>Trong Hieu Tran and Ngoc Thanh Nguyen</i>	

Knowledge Based and Expert Systems

Emotion Judgment Based on Relationship between Speaker and Sentential Actor	62
<i>Seiji Tsuchiya, Eriko Yoshimura, Fuji Ren, and Hirokazu Watabe</i>	
A Knowledge Based Formal Language for Securing Information Systems	70
<i>Yun Bai</i>	
Multi Criteria Decision Making in Fuzzy Description Logics: A First Step	78
<i>Umberto Straccia</i>	

A Hybrid System Combining Description Logics and Rules for Inventive Design	87
<i>Alexis Bultey, Cecilia Zanni-Merk, François Rousselot, and François de Beuvron</i>	
Domain Modeling Based on Engineering Standards	95
<i>Carlos Toro, Manuel Graña, Jorge Posada, Javier Vaquero, Cesar Sanín, and Edward Szczerbicki</i>	
A Knowledge Based System for Minimum Rectangle Nesting	103
<i>Grzegorz Chmaj, Iwona Pozniak-Koszalka, and Andrzej Kasprzak</i>	

Other/Misc. Generic Intelligent Systems Topics

Evolutionary Algorithm for Solving Congestion Problem in Computer Networks	112
<i>Dawid Ohia, Leszek Koszalka, and Andrzej Kasprzak</i>	
Automatic Speech-Lip Synchronization System for 3D Animation	122
<i>Juan Monroy, Francisco Bellas, Richard J. Duro, Ruben Lopez, Antonio Puentes, and Jacques Isaac</i>	
Development of an Effective Travel Time Prediction Method Using Modified Moving Average Approach	130
<i>Nihad Karim Chowdhury, Rudra Pratap Deb Nath, Hyunjo Lee, and Jaewoo Chang</i>	
Differential Evolution and Genetic Algorithms for the Linear Ordering Problem	139
<i>Václav Snášel, Pavel Krömer, and Jan Platoš</i>	
Determining Optimal Crop Rotations by Using Multiobjective Evolutionary Algorithms	147
<i>Ruth Pavón, Ricardo Brunelli, and Christian von Lücken</i>	

Intelligent Vision and Image Processing

Object Recognition by Permanence of Ratios Based Fusion and Gaussian Bayes Decision	155
<i>Tuan D. Pham</i>	
A New Wavelet–Fractal Image Compression Method	161
<i>Vu Thanh Hien</i>	
Urban Vehicle Tracking Using a Combined 3D Model Detector and Classifier	169
<i>Norbert Buch, Fei Yin, James Orwell, Dimitrios Makris, and Sergio A. Velastin</i>	

UBIAS – Type Cognitive Systems for Medical Pattern Interpretation . . .	177
<i>Lidia Ogiela, Marek R. Ogiela, and Ryszard Tadeusiewicz</i>	
A New Selective Confidence Measure–Based Approach for Stereo Matching	184
<i>Nizar Fakhfakh, Louahdi Khoudour, El-Miloudi El-Koursi, Jacques Jacot, and Alain Dufaux</i>	
Image Content Analysis for Cardiac 3D Visualizations	192
<i>Miroslaw Trzupek, Marek R. Ogiela, and Ryszard Tadeusiewicz</i>	
Knowledge Management, Ontologies and Data Mining	
Illogical Adjective Phrase Detection for Computer Conversation	200
<i>Eriko Yoshimura, Seiji Tsuchiya, Hirokazu Watabe, and Tsukasa Kawaoka</i>	
A Non-sequential Representation of Sequential Data for Churn Prediction	209
<i>Mark Eastwood and Bogdan Gabrys</i>	
Dialectics-Based Knowledge Acquisition – A Case Study	219
<i>Cecilia Zanni-Merk and Philippe Bouché</i>	
Automatic Extraction of Hyponymy-Hypernymy Lexical Relations between Nouns from a Spanish Dictionary	227
<i>Rodolfo A. Pazos R., José A. Martínez F., Juan J. González B., María Lucila Morales-Rodríguez, and Jessica C. Rojas P.</i>	
AVEDA: Statistical Tests for Finding Interesting Visualisations	235
<i>Katharina Tschumitschew and Frank Klawonn</i>	
Degree of Association between Documents Using Association Mechanism	243
<i>Hirokazu Watabe, Eriko Yoshimura, and Seiji Tsuchiya</i>	
Parallel Method for Mining High Utility Itemsets from Vertically Partitioned Distributed Databases	251
<i>Bay Vo, Huy Nguyen, Tu Bao Ho, and Bac Le</i>	
An Ontology-Based Autonomic System for Improving Data Warehouse Performances	261
<i>Vlad Nicoliciu-Georgescu, Vincent Benatier, Remi Lehn, and Henri Briand</i>	
Semantic Enhancement of the Course Curriculum Design Process	269
<i>Javier Vaquero, Carlos Toro, Juantru Martín, and Andoni Aregita</i>	

Using the Mesh Thesaurus to Index a Medical Article: Combination of Content, Structure and Semantics 277
Jihen Majdoubi, Mohamed Tmar, and Faiez Gargouri

Web Intelligence, Text and Multimedia Mining and Retrieval

Building Decision Trees to Identify the Intent of a User Query 285
Marcelo Mendoza and Juan Zamora

Ontology-Based Concept Indexing of Images 293
Rossitza Setchi, Qiao Tang, and Carole Bouchard

Design and Implementation of a Methodology for Identifying Website Keyobjects 301
Luis E. Dujovne and Juan D. Velásquez

NLP Contribution to the Semantic Web: Linking the Term to the Concept 309
Gaëlle Lortal, Nathalie Chaignaud, Jean-Philippe Kotowicz, and Jean-Pierre Pécuchet

An Intelligent Automatic Hoax Detection System 318
Marin Vuković, Krešimir Pripuzić, and Hrvoje Belani

Web User Session Reconstruction with Back Button Browsing 326
Robert F. Dell, Pablo E. Román, and Juan D. Velásquez

Other Advanced Knowledge-Based Systems (I)

Fast Time Delay Neural Networks for Detecting DNA Coding Regions 333
Hazem M. El-Bakry and Mohamed Hamada

Consistency-Based Feature Selection 342
Kilho Shin and Xian Ming Xu

Asynchronous Situated Coevolution and Embryonic Reproduction as a Means to Autonomously Coordinate Robot Teams 351
Abraham Prieto, Francisco Bellas, Andres Faina, and Richard J. Duro

Keynote Speaker Plenary Presentation

Learning Automata Based Intelligent Tutorial-like System 360
B. John Oommen and M. Khaled Hashem

Modeling and Simulating Empires: Toward a Game World Generator (Abstract)	374
<i>Barry G. Silverman</i>	
User-Centric and Intelligent Service Composition in Ubiquitous Computing Environments (Abstract)	375
<i>In-Young Ko</i>	
Author Index	377

Table of Contents – Part II

Innovations in Chance Discovery

Discourse Analysis of Communication Generating Social Creativity	1
<i>Yoko Nishihara, Yuichi Takahashi, and Yukio Ohsawa</i>	
Value Cognition System as Generalization of Chance Discovery	9
<i>Yukio Ohsawa</i>	
Temporal Logic for Modeling Discovery and Logical Uncertainty	16
<i>Sergey Babenyshev and Vladimir V. Rybakov</i>	
Evaluation of a Classification Rule Mining Algorithm Based on Secondary Differences	24
<i>Shusaku Tsumoto and Hidenao Abe</i>	
Communication between Living and Scientific Knowledge as Chance Discovery	32
<i>Yumiko Nara</i>	

Advanced Knowledge-Based Systems

Automatically Estimating and Updating Input-Output Tables	42
<i>Ting Yu, Manfred Lenzen, Chris Dey, and Jeremy Badcock</i>	
Context-Aware User and Service Profiling by Means of Generalized Association Rules	50
<i>Elena Baralis, Luca Cagliero, Tania Cerquitelli, Paolo Garza, and Marco Marchetti</i>	
An ETL Tool Based on Semantic Analysis of Schemata and Instances	58
<i>Sonia Bergamaschi, Francesco Guerra, Mirko Orsini, Claudio Sartori, and Maurizio Vincini</i>	
Knowledge Source Discovery: An Experience Using Ontologies, WordNet and Artificial Neural Networks	66
<i>Mariano Rubiolo, María Laura Caliusco, Georgina Stegmayer, Matías Gareli, and Mauricio Coronel</i>	
Path Planning Knowledge Modeling for a Generic Autonomous Robot: A Case Study	74
<i>Rafael Guirado, Clara Marcela Miranda, and José Fernando Bienvenido</i>	

System Models for Goal-Driven Self-management in Autonomic Databases 82
Marc Holze and Norbert Ritter

\mathcal{I} -SQE: A Query Engine for Answering Range Queries over Incomplete Spatial Databases 91
Alfredo Cuzzocrea and Andrea Nucita

Multi-Agent Negotiation and Coordination: Models and Applications

An Agent-Mediated Collaborative Negotiation in E-Commerce: A Case Study in Travel Industry 102
Bala M. Balachandran, Ebrahim Alhashel, and Masoud Mohammedian

The Role of Ontology in Modelling Autonomous Agent-Based Systems 111
Ebrahim Alhashel, Bala M. Balachandran, and Dharmendra Sharma

Multi-Agent Systems in Quantum Security for Modern Wireless Networks 119
Xu Huang and Dharmendra Sharma

Innovations in Intelligent Systems (I)

Tolerance Classes in Measuring Image Resemblance 127
A.H. Meghdadi, J.F. Peters, and S. Ramanna

Capillary Blood Vessel Tortuosity Measurement Using Graph Analysis 135
Mariusz Paradowski, Halina Kwasnicka, and Krzysztof Borysewicz

Image Features Based on Local Hough Transforms 143
Andrzej Śluzek

Capillary Abnormalities Detection Using Vessel Thickness and Curvature Analysis 151
Mariusz Paradowski, Urszula Markowska-Kaczmar, Halina Kwasnicka, and Krzysztof Borysewicz

Intelligent Technology Approach to Management Engineering

A Hybrid Method of Biological Computation and Genetic Algorithms for Resolving Process-Focused Scheduling Problems 159
Ikno Kim and Junzo Watada

Searching Cliques in a Fuzzy Graph Based on an Evolutionary and Biological Method	166
<i>Ikno Kim and Junzo Watada</i>	
A Biologically Intelligent Encoding Approach to a Hierarchical Classification of Relational Elements in a Digraph	174
<i>Ikno Kim and Junzo Watada</i>	
A Bio-inspired Evolutionary Approach to Identifying Minimal Length Decision Rules in Emotional Usability Engineering	181
<i>Ikno Kim and Junzo Watada</i>	
Determining Workstation Groups in a Fixed Factory Facility Based on Biological Computation	188
<i>Ikno Kim and Junzo Watada</i>	
A Fuzzy Risk Assessment in Software Development Defuzzified by Signed Distance	195
<i>Huey-Ming Lee and Lily Lin</i>	
Particle Swarm Optimization for Multi-function Worker Assignment Problem	203
<i>Shamshul Bahar Yaakob and Junzo Watada</i>	
Evidential Reasoning Based on DNA Computation	212
<i>Rohani Binti Abu Bakar and Junzo Watada</i>	
Dynamic Tracking System through PSO and Parzen Particle Filter	220
<i>Zalili Binti Musa, Junzo Watada, Sun Yan, and Haochen Ding</i>	
 Data Mining and Service Science for Innovation 	
Text Mining for Customer Enquiries in Telecommunication Services	228
<i>Motoi Iwashita, Shinsuke Shimogawa, and Ken Nishimatsu</i>	
Defuzzification Using Area Method on L^∞ Space	236
<i>Takashi Mitsuishi and Yasunari Shidama</i>	
Agent-Based In-Store Simulator for Analyzing Customer Behaviors in a Super-Market	244
<i>Takao Terano, Ariyuki Kishimoto, Toru Takahashi, Takashi Yamada, and Masakazu Takahashi</i>	
Detecting Temporal Patterns of Importance Indices about Technical Phrases	252
<i>Hidenao Abe and Shusaku Tsumoto</i>	
Recommender System for Music CDs Using a Graph Partitioning Method	259
<i>Takanobu Nakahara and Hiroyuki Morita</i>	

Optimization of Budget Allocation for TV Advertising	270
<i>Kohei Ichikawa, Katsutoshi Yada, Namiko Nakachi, and Takashi Washio</i>	

Knowledge-Based Systems for e-Business

Cover All Query Diffusion Strategy over Unstructured Overlay Network	278
<i>Yoshikatsu Fujita, Yasufumi Saruwatari, Masakazu Takahashi, and Kazuhiko Tsuda</i>	
Extracting the Potential Sales Items from the Trend Leaders with the ID-POS Data	285
<i>Masakazu Takahashi, Kazuhiko Tsuda, and Takao Terano</i>	
A Study on Comprehending the Intention of Administrative Documents in the Field of e-Government	293
<i>Keiichiro Mitani, Yoshinori Fukue, and Kazuhiko Tsuda</i>	
Decision Making Process for Selecting Outsourcing Company Based on Knowledge Database	300
<i>Akihiro Hayashi, Yasunobu Kino, and Kazuhiko Tsuda</i>	
Intelligent QA Systems Using Semantic Expressions	308
<i>Yutaka Inada, Hideo Nakano, Shinkaku Kashiji, and Junichi Aoe</i>	
The Effective Extraction Method for the Gap of the Mutual Understanding Based on the Egocentrism in Business Communications	317
<i>Nobuo Suzuki and Kazuhiko Tsuda</i>	

Innovations in Intelligent Systems (II)

Deriving Electrical Dependencies from Circuit Topologies Using Logic Grammar	325
<i>Takushi Tanaka</i>	
A Fast Nearest Neighbor Method Using Empirical Marginal Distribution	333
<i>Mineichi Kudo, Jun Toyama, and Hideyuki Imai</i>	
Effects of Kurtosis for the Error Rate Estimators Using Resampling Methods in Two Class Discrimination	340
<i>Kozo Yamada, Hirohito Sakurai, Hideyuki Imai, and Yoshiharu Sato</i>	
Reasoning about External Environment from Web Sources	348
<i>Hércules Antonio do Prado, André Ribeiro Magalhães, and Edilson Ferneda</i>	

An Agent Control Method Based on Variable Neighborhoods	356
<i>Seiki Ubukata, Yasuo Kudo, and Tetsuya Murai</i>	
<i>Counselor</i> , a Data Mining Based Time Estimation for Software Maintenance	364
<i>Hércules Antonio do Prado, Edilson Ferneda, Nicolas Anquetil, and Elizabeth d'Arrochella Teixeira</i>	
An Integrated Knowledge Adaption Framework for Case-Based Reasoning Systems	372
<i>Ning Lu, Jie Lu, and Guangquan Zhang</i>	
A Logical Anticipatory System of Before-After Relation Based on Bf-EVALPSN	380
<i>Kazumi Nakamatsu, Jair Minoro Abe, and Seiki Akama</i>	
A Note on Monadic Curry System P_1	388
<i>Jair Minoro Abe, Kazumi Nakamatsu, and Fábio Romeu de Carvalho</i>	

Video Surveillance

Adaptation of Space-Mapping Methods for Object Location Estimation to Camera Setup Changes — A New Study	395
<i>Chih-Jen Wu and Wen-Hsiang Tsai</i>	
A Novel Method for Lateral Vehicle Localization by Omni-Cameras for Car Driving Assistance	403
<i>Chih-Jen Wu and Wen-Hsiang Tsai</i>	
Abnormal Event Analysis Using Patching Matching and Concentric Features	411
<i>Jun-Wei Hsieh, Sin-Yu Chen, and Chao-Hong Chiang</i>	
Video Inpainting on Digitized Old Films	421
<i>Nick C. Tang, Hong-Yuan Mark Liao, Chih-Wen Su, Fay Huang, and Timothy K. Shih</i>	
Pedestrian Identification with Distance Transform and Hierarchical Search Tree	431
<i>Daw-Tung Lin and Li-Wei Liu</i>	
An Enhanced Layer Embedded Pedestrian Detector	439
<i>Duan-Yu Chen</i>	

Social Networks

Mining Influential Bloggers: From General to Domain Specific	447
<i>Yichuan Caiv and Yi Chen</i>	

Efficiency of Node Position Calculation in Social Networks	455
<i>Piotr Brodka, Katarzyna Musiał, and Przemysław Kazienko</i>	
Time Series Analysis of R&D Team Using Patent Information	464
<i>Yurie Ino and Sachio Hirokawa</i>	
Extracting Research Communities by Improved Maximum Flow Algorithm.	472
<i>Toshihiko Horŭke, Youhei Takahashi, Tetsuji Kuboyama, and Hiroshi Sakamoto</i>	
Virtual Communities of Practice’s Purpose Evolution Analysis Using a Concept-Based Mining Approach	480
<i>Sebastián A. Ríos, Felipe Aguilera, and Luis A. Guerrero</i>	
Discovering Networks for Global Propagation of Influenza A (H3N2) Viruses by Clustering	490
<i>Kazuya Sata, Kouichi Hirata, Kimihito Ito, and Tetsuji Kuboyama</i>	
Advanced Engineering Design Techniques for Adaptive Systems	
Machine Vision Application to Automatic Intruder Detection Using CCTV.	498
<i>Hernando Fernandez-Canque, Sorin Hintea, John Freer, and Ali Ahmadinia</i>	
A Genetic Algorithm-Based Multiobjective Optimization for Analog Circuit Design	506
<i>Gabriel Oltean, Sorin Hintea, and Emilia Sipos</i>	
Optimization of Reconfigurable Multi-core SOCs for Multi-standard Applications.	515
<i>Ali Ahmadinia, Tughrul Arslan, and Hernando Fernandez Canque</i>	
Knowledge Technology in Learning Support	
Vocabulary Learning Environment with Collaborative Filtering for Support of Self-regulated Learning	523
<i>Masanori Yamada, Satoshi Kitamura, Shiori Miyahara, and Yuhei Yamauchi</i>	
Online Collaboration Support Tools for Project-Based Learning of Embedded Software Design	531
<i>Takashi Yukawa, Hirotaka Takahashi, Yoshimi Fukumura, Makoto Yamazaki, Toshimasa Miyazaki, Shohei Yano, Akiko Takeuchi, Hajime Miura, and Naoki Hasegawa</i>	

The Relationship between the Learning Styles of the Students and Their e-Learning Course Adaptability	539
<i>Kazunori Nishino, Hiroko Toya, Shinji Mizuno, Kumiko Aoki, and Yoshimi Fukumura</i>	
Effectiveness of Engineering Solution Case Document Search Based on TRIZ Contradiction Matrix Theory	547
<i>Koji Yamada, Motoki Miura, Tessai Hayama, and Susumu Kunifuji</i>	
A Following Method of Annotations on Updated Contents and Its Evaluation	555
<i>Hisayoshi Kunimune, Kenzou Yokoyama, Takeshi Takizawa, and Yasushi Fuwa</i>	
Organization of Solution Knowledge Graph from Collaborative Learning Records	564
<i>Yuki Watanabe, Tomoko Kojiri, and Toyohide Watanabe</i>	
Implementation of Wireless Sensor System and Interface for Agricultural Use	572
<i>Kenji Obata, Takahiro Masui, Hiroshi Mineno, and Tadanori Mizuno</i>	
Algorithms for Extracting Topic across Different Types of Documents	580
<i>Shoichi Nakamura, Saori Chiba, Hirokazu Shirai, Hiroaki Kaminaga, Setsuo Yokoyama, and Youzou Miyadera</i>	
Advanced Information System for Supporting Personal Activity	
Face Image Annotation in Impressive Words by Integrating Latent Semantic Spaces and Rules	591
<i>Hideaki Ito, Yuji Kawai, and Hiroyasu Koshimizu</i>	
Sketch Learning Environment for Human Body Figure by Imitative Drawing	599
<i>Masato Soga, Takahisa Fukuda, and Hirokazu Taki</i>	
Design and Implementation of an Optimal Radio Access Network Selection Algorithm Using Mutually Connected Neural Networks	607
<i>Mikio Hasegawa, Taichi Takeda, and Hiroshi Harada</i>	
Probabilistic Estimation of Travel Behaviors Using Zone Characteristics	615
<i>Masatoshi Takamiya, Kosuke Yamamoto, and Toyohide Watanabe</i>	
A Web-Based Approach for Automatic Composition of an Insightful Slideshow for Personal Photographs	623
<i>Kotaro Yatsugi, Naomi Fujimura, and Taketoshi Ushiamo</i>	

Design of Intelligent Society

Web-Based System for Supporting Participation in International Conferences	631
<i>Akira Hattori, Shigenori Irooi, and Haruo Hayami</i>	
Analyzing the Relationship between Complexity of Road Networks and Mobile Agents' Simulation	639
<i>Kazunori Iwata, Nobuhiro Ito, Yoichi Setoguchi, and Naohiro Ishii</i>	
Multi-base Station Placement for Wireless Reprogramming in Sensor Networks	648
<i>Aoi Hashizume, Hiroshi Mineno, and Tadanori Mizuno</i>	
Optimization of Transport Plan for On-Demand Bus System Using Electrical Vehicles	656
<i>Kousuke Kawamura and Naoto Mukai</i>	
Public Large Screen Enabled Content Collection and Connection	664
<i>Kosuke Numa, Hironori Tomobe, Tatsuo Sugimoto, Masako Miyata, Kiyoko Toriumi, Jun Abe, and Koichi Hori</i>	

Knowledge-Based Interface Systems (I)

Implementing Multi-relational Mining with Relational Database Systems	672
<i>Nobuhiro Inuzuka and Toshiyuki Makino</i>	
A Simple Method for 3-Dimensional Photorealistic Facial Modeling and Consideration the Reconstructing Error	681
<i>Ippei Torii, Yousuke Okada, Masayuki Mizutani, and Naohiro Ishii</i>	
Study of Writer Recognition by Japanese Hiragana	689
<i>Yoshinori Adachi, Masahiro Ozaki, and Yuji Iwahori</i>	

Knowledge-Based Interface Systems (II)

Speed Flexibility Biomedical Vision Model Using Analog Electronic Circuits and VLSI Layout Design	697
<i>Masashi Kawaguchi, Shoji Suzuki, Takashi Jimbo, and Naohiro Ishii</i>	
Self-calibration and Image Rendering Using RBF Neural Network	705
<i>Yi Ding, Yuji Iwahori, Tsuyoshi Nakamura, Robert J. Woodham, Lifeng He, and Hidenori Itoh</i>	
Similarity Grouping of Paintings by Distance Measure and Self Organizing Map	713
<i>Naohiro Ishii, Yusaku Tokuda, Ippei Torii, and Tomomi Kanda</i>	

Knowledge-Based Multi-Criteria Decision Support

Localization in Wireless Sensor Networks by Fuzzy Logic System	721
<i>Shu-Yin Chiang and Jin-Long Wang</i>	
Dynamic Handover Scheme for WiMAX	729
<i>Jin-Long Wang and Shu-Yin Chiang</i>	
A Fuzzy Bilevel Model and a PSO-Based Algorithm for Day-Ahead Electricity Market Strategy Making	736
<i>Guangquan Zhang, Guoli Zhang, Ya Gao, and Jie Lu</i>	
Correspondence between Incomplete Fuzzy Preference Relation and Its Priority Vector	745
<i>Pei-Di Shen, Wen-Li Chyr, Hsuan-Shih Lee, and Kuang Lin</i>	

Soft Computing Techniques and Their Applications

Nature Inspired Design of Autonomous Driving Agent—Realtime Localization, Mapping and Avoidance of Obstacle Based on Motion Parallax	752
<i>Ivan Tanev and Katsunori Shimohara</i>	
Effective Utilization of Neural Networks for Constructing an Intelligent Decision Support System for Dealing Stocks	761
<i>Norio Baba and Kou Nin</i>	
Fine Grained Parallel Processing for Soft Computing	767
<i>Osamu Fujita and Koji Jinya</i>	
New System Structuring Method That Adapts to Technological Progress of Semiconductors	773
<i>Kunihiko Yamada, Kouji Yoshida, Masanori Kojima, Tetuya Matumura, and Tadanori Mizuno</i>	

Immunity-Based Systems

A Network Approach for HIV-1 Drug Resistance Prevention	782
<i>Kouji Harada and Yoshiteru Ishida</i>	
Asymmetric Phenomena of Segregation and Integration in Biological Systems: A Matching Automaton	789
<i>Yoshiteru Ishida and Tatsuya Hayashi</i>	
Adaptive Forecasting of High-Energy Electron Flux at Geostationary Orbit Using ADALINE Neural Network	797
<i>Masahiro Tokumitsu, Yoshiteru Ishida, Shinichi Watari, and Kentarou Kitamura</i>	

A Note on Biological Closure and Openness: A System Reliability
View 805
Yoshiteru Ishida

Other Advanced Knowledge-Based Systems (II)

Faith in the Algorithm, Part 2: Computational Eudaemonics 813
Marko A. Rodriguez and Jennifer H. Watkins

System Engineering Security 821
Esmiralda Moradian

A Semantically-Based Task Model and Selection Mechanism in
Ubiquitous Computing Environments 829
*Angel Jimenez-Molina, Jun-Sung Kim, Hyung-Min Koo,
Byung-Seok Kang, and In-Young Ko*

A Platform for Extracting and Storing Web Data 838
L. Víctor Rebolledo and Juan D. Velásquez

Bayesian Reflectance Component Separation 846
*Ramón Moreno, Manuel Graña, Alicia d'Anjou, and
Carmen Hernandez*

Identifying Fewer Key Factors by Attribute Selection Methodologies to
Understand the Hospital Admission Prediction Pattern with Ant Miner
and C4.5 853
Kyoko Fukuda

Combined Unsupervised-Supervised Classification Method 861
Urszula Markowska-Kaczmar and Tomasz Switek

Author Index 869

Intersection Search for a Fuzzy Petri Net-Based Knowledge Representation Scheme

Slobodan Ribarić¹, Nikola Pavešić², and Valentina Zadrija¹

¹ Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia

² Faculty of Electrical Engineering, University of Ljubljana, Slovenia

{slobodan.ribaric, valentina.zadrija}@fer.hr
nikola.pavesic@fe.uni-lj.si

Abstract. This paper describes the intersection search as an inference procedure for a knowledge representation scheme based on the theory of Fuzzy Petri Nets. The procedure uses the dynamical properties of the scheme. The relationships between the concepts of interest, obtained by the intersection search algorithm, are accompanied by the value of the linguistic variable expressing the assurance for the relations. An illustrative example of the intersection search procedure is provided.

Keywords: Knowledge representation, Inference procedure, Fuzzy Petri Net, Intersection search.

1 Introduction

One of the central problems of the knowledge-based systems, especially for real-world tasks solving, where knowledge is based on vague, uncertain and fuzzy facts, is the development of a knowledge representation scheme. Among the knowledge representation schemes that support uncertain and fuzzy knowledge representation and reasoning there is a class of schemes based on the theory of Fuzzy Petri Nets (FPNs) [1]: Looney [2] and Chen et al. [3] proposed FPNs for rule-based decision making; Scarpelli et al. [4] described a reasoning algorithm for a high-level FPN; Chen [5] introduced a Weight FPN model for rule-based systems; Li and Lara-Rosano [6] proposed a model based on an Adaptive FPN, which is implemented for knowledge inference; Looney and Liang [7] proposed the fuzzy-belief Petri Nets (PN) as combination of the bi-directional fuzzy propagation of the fuzzy-belief network and the FPN; Lee et al. [8] introduced a reasoning algorithm based on possibilistic PN as a mechanism that mimics human inference; Canales et al. [9] described a method of fuzzy-knowledge learning based on an Adaptive FPN; Ha et al. [10] described knowledge representation by weighted fuzzy-production rules and inference with a generalized FPN; and Guo-Yan [11] proposed a hybrid of the PN and the Fuzzy PN to support an inference procedure. Shen [12] presented a knowledge-representation scheme based on a high-level FPN for modeling fuzzy IF-THEN-ELSE rules.

In this paper the intersection search procedure based on “spreading activation” for a Fuzzy Petri Net-based knowledge representation scheme is proposed.

2 A Fuzzy Petri Net-Based Knowledge Representation Scheme

A network-based fuzzy knowledge representation scheme named KRFPN (**K**nowledge-**R**epresentation Scheme based on the **F**uzzy **P**etri-**N**ets theory) uses the concepts of the Fuzzy Petri Net theory to represent uncertain, vague and/or fuzzy information obtained from modeled, real-world situations. The knowledge representation scheme KRFPN is defined as the 13-tuple:

$$\text{KRFPN} = (P, T, I, O, M, \Omega, \mu, f, c, \lambda, \alpha, \beta, C), \quad (1)$$

where the first 10 components represent a Fuzzy Petri net FPN [18] defined as follows: $P = \{p_1, p_2, \dots, p_n\}$ is a finite set of places; $T = \{t_1, t_2, \dots, t_m\}$ is a finite set of transitions; $P \cap T = \emptyset$; $I: T \rightarrow P^\infty$ is an input function, a mapping from transitions to bags of places; $O: T \rightarrow P^\infty$ is an output function, a mapping from transitions to bags of places; $M = \{m_1, m_2, \dots, m_r\}$, $1 \leq r < \infty$, is a set of tokens; $\Omega: P \rightarrow \wp(M)$ is a mapping, from P to $\wp(M)$, called a distribution of tokens, where $\wp(M)$ denotes the power set of M . Using Ω_0 we denote the initial distribution of tokens in the places of an FPN; $\mu: P \rightarrow \mathbb{N}$ is a marking, a mapping from places to non-negative integers, \mathbb{N} . A mapping μ can be represented as an n -component vector $\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_n)$, where n is a cardinality of the set P . An initial marking is denoted by the vector $\boldsymbol{\mu}_0$. A function $f: T \rightarrow [0, 1]$ is an association function, a mapping from transitions to real values between zero and one, and $c: M \rightarrow [0, 1]$ is an association function, a mapping from tokens to real values between zero and one, and $\lambda \in [0, 1]$ is a threshold value related to the firing of a FPN.

A marked FPN can be represented by a bipartite directed multi-graph containing two types of nodes: places (graphically represented by circles) and transitions (bars). The relationships that are based on input and output functions are represented by directed arcs. Each arc is directed from an element of one set (P or T) to the element of another set (T or P). The tokens in the marked FPN graphs are represented by labeled dots $c(m_l)$, where $c(m_l)$ denotes the value of the token.

Tokens give dynamical properties to an FPN, and they are used to define its *execution*, i.e., by firing an enabled transition t_j , tokens are removed from its input places (elements in $I(t_j)$). Simultaneously, new tokens are created and distributed to its output places (elements of $O(t_j)$). In an FPN, a transition t_j is *enabled* if each of its input places has at least as many tokens in it as there are arcs from the place to the transition and if the values of the tokens $c(m_l)$, $l = 1, 2, \dots$ exceed a threshold value $\lambda \in [0, 1]$. The number of tokens at the input and output places of the fired transition is changed in accordance with the basic definition of the original PN [16]. The new token value in the output place is $c(m_l)f(t_j)$, where $c(m_l)$ is the value of the token at the input place $p_j \in I(t_j)$ and $f(t_j)$ is the degree of truth of the relation assigned to the transition $t_j \in T$.

The components α , β and C , introduce a semantic interpretation to the scheme: $\alpha: P \rightarrow D$ is a bijective function that maps a set of places onto a set of concepts D . The set of concepts D consists of the formal objects used for representing objects and facts from the agent's world. The elements from $D = D_1 \cup D_2 \cup D_3$ are as follows: elements that denote classes or categories of objects and represent higher levels of abstraction (D_1), elements corresponding to individual objects as the instances of the

classes (D_2) and those elements representing the intrinsic properties of the concepts or values of these properties (D_3).

$\beta: T \rightarrow \Sigma$ is a surjective function that associates a description of the relationship among the facts and objects to every transition $t_i \in T$; $i = 1, 2, \dots, m$, where m is a cardinality of the set T . The set $\Sigma = \Sigma_1 \cup \Sigma_2 \cup \Sigma_3$ consists of elements corresponding to the relationships between the concepts used for the partial ordering of the set of concepts (Σ_1), the elements used to specify the types of properties to which the values from subset D_3 are assigned (Σ_2), and the elements corresponding to the relationships between the concepts, but not used for hierarchical structuring (Σ_3). For example, elements from Σ_3 may be used to specify the spatial relations among the objects.

The semantic interpretation requires the introduction of a set of contradictions C , which is a set of pairs of mutually contradictory relations (for example, *is_a* and *is_not_a*), as well as, pairs of mutually contradictory concepts if they are inherited for the same concept or object (for example, the object cannot simultaneously inherit properties such as “*Quadruped*” and “*Biped*”).

The inverse function $\alpha^{-1}: D \rightarrow P$, and the generalized inverse function $\beta^{-1}: \Sigma \rightarrow \tau$; $\tau \subseteq T$ are defined in the KRFPN.

Note that the KRFPN inherits dynamical properties from the FPN.

The uncertainty and confidence related to the facts, concepts and the relationships between them in the KRFPN are expressed by means of the values of the $f(t_i)$, $t_i \in T$, and $c(m_i)$, $m_i \in M$, association functions. The value of the function f , as well as the value of the function c , can be expressed by the truth scales and by their corresponding numerical intervals proposed in [3] - from “*always true*” [1.0, 1.0], “*extremely true*” [0.95, 0.99], “*very true*” [0.80, 0.94] to “*minimally true*” [0.01, 0.09], and “*not true*” [0.0, 0.0].

Example 1

In order to illustrate the basic components of the KRFPN, a simple example of the agent’s knowledge base for a scene (adapted from [17]; Fig. 1) is introduced. Briefly, the scene may be described as follows: Shaggy, who is a human, and Scooby, the dog, are cartoon characters. Scooby, as a cartoon character, can talk and he is, like Shaggy, a mammal. Shaggy wears clothes and he is in front of Scooby. We suppose that both Shaggy and Scooby are hungry.

The knowledge base designed by the KRFPN has the following components (Fig. 2):

$$\begin{aligned} P &= \{p_1, p_2, \dots, p_{12}\}; T = \{t_1, t_2, \dots, t_{17}\}; \\ I(t_1) &= \{p_1\}; I(t_2) = \{p_3\}, \dots; I(t_{17}) = \{p_1\}; \\ O(t_1) &= \{p_5\}; O(t_2) = \{p_4\}, \dots; O(t_{17}) = \{p_3\}. \end{aligned}$$

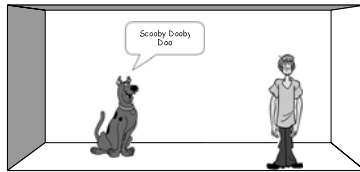


Fig. 1. A simple scene with *Scooby* and *Shaggy* [17]

The initial distribution of tokens is:

$\Omega_0 = \{\{m_1\}, \emptyset, \dots, \emptyset\}$, where $c(m_1) = 1.0$, where \emptyset denotes an empty set.

The vector $\mu_0 = (1, 0, \dots, 0)$ denotes that there is only one token in the place p_1 . The function f is specified as follows:

$f(t_1) = f(t_2) = \dots = f(t_9) = 1.0$; $f(t_{10}) = f(t_{12}) = 0.8$; and $f(t_{11}) = \dots = f(t_{17}) = 0.9$; $f(t_i)$, $i = 1, 2, \dots, m$ indicates the degree of our pursuance in the truth of the relation $\beta(t_i)$.

The set D is defined as follows:

$D_1 = \{\text{Cartoon_Character, Cartoon_Dog, Human, Mammal, Dog, Live}\}$, $D_2 = \{\text{Scooby, Shaggy}\}$ and $D_3 = \{\text{Talking, Brown, Hungry, Wears_Clothes}\}$.

The set Σ consists of:

$\Sigma_1 = \{\text{is_a, is_not_a}\}$, $\Sigma_2 = \{\text{has_characteristic, has_not_characteristic, has_color}\}$ and $\Sigma_3 = \{\text{is_in_front_of, is_behind_of}\}$.

Functions α and β are: $\alpha: p_1 \rightarrow \text{Shaggy}$, $\beta: t_1 \rightarrow \text{is_a}$, $\alpha: p_2 \rightarrow \text{Cartoon_Character}$, $\beta: t_2 \rightarrow \text{is_a}$, ..., $\alpha: p_{12} \rightarrow \text{Wears_Clothes}$, $\beta: t_{17} \rightarrow \text{is_in_front_of}$.

A set of contradictions C is $\{\{\text{has_characteristic, has_not_characteristic}\}, \{\text{is_in_front_of, is_behind_of}\}, \{\text{is_a, is_not_a}\}\}$.

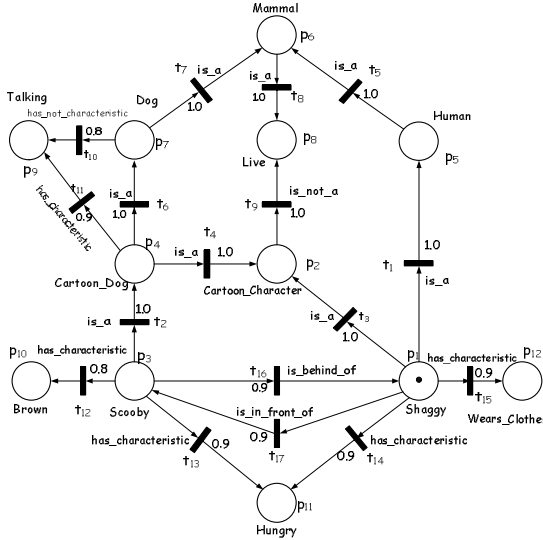


Fig. 2. The knowledge base designed by the KRFPN (Example 1)

For the initial distribution of tokens, the following transitions are enabled: t_1, t_3, t_{14}, t_{15} and t_{17} .

3 Intersection Search Algorithm

R. Quillian proposed a procedure that corresponds to the inference in semantic networks [14]. The procedure, called *intersection search* or *spreading activation*, makes

it possible to find relationships between facts stored in a knowledge base by “spreading activation” out of two nodes (called *patriarch nodes*) and finding their intersection. The nodes where the activations meet are called *intersection nodes*. The paths from two patriarch nodes to the intersection nodes define the relationships between the facts.

Based on the above idea the intersection search algorithm for the fuzzy knowledge representation scheme KRFPN is here proposed. The intersection search inference procedure in the KRFPN is based on its dynamical properties, given by the firing enabled transitions, and the determination of the inheritance set of the KRFPN. The inheritance set for the KRFPN is based on concepts similar to the reachability set of the ordinary Petri nets (PNs), where the reachability relationship is the reflexive, transitive closure of the immediately reachable relationship [16]. The reachability set of the PN is graphically represented by a *reachability tree*.

The main differences between the inheritance set of the KRFPN and the reachability set of the PN [16] are as follows: (i) After firing an enabled transition, where the transition is related to the element in the subset Σ_1 (recall that the elements in Σ_1 are used for the hierarchical structuring) that specifies an exception or negation (for example, *is_not_a*), the created token(s) in the corresponding output place(s) has to be frozen. A frozen token in the output place is fixed and it cannot enable a transition. (ii) After firing all the enabled transitions for the distribution of tokens in the KRFPN, where the transitions are related to the elements in the subsets Σ_2 and Σ_3 , the created tokens at the corresponding output places also have to be *frozen*. Recall that the elements in Σ_2 and Σ_3 are used to specify the properties and the non-hierarchical structuring, respectively. (iii) An inheritance tree, as a graphical representation of the inheritance set, is bounded by $k + 1$ levels, where k is a predefined number of levels. Such an inheritance tree is called a k -level inheritance tree. (iv) A k -level inheritance tree has the following additional types of nodes: a k -terminal node, a frozen node, and an identical node.

Taking into account the above particularities, a k -level inheritance tree can be constructed by applying the slightly modified algorithm for the reachability tree given in [16]. The algorithm for construction of a k -level inheritance tree is given in [18].

A k -level inheritance tree consists of the nodes $\boldsymbol{\pi} = (\pi_1, \pi_2, \dots, \pi_n)$, where n is the cardinality of the set of places P , and the directed, labeled arcs. In order to simplify and make the notation uniform, the nodes in the tree are denoted by n -component vectors in the form $\boldsymbol{\pi} = (\pi_1, \pi_2, \dots, \pi_n)$. Each component π_i ; $i = 1, 2, \dots, n$ of $\boldsymbol{\pi}$ is represented by an empty set \emptyset for $\mu(p_i) = 0$, i.e., there is no token(s) at the place p_i , or by a set $\{c(m_k), \dots, c(m_l), \dots, c(m_s)\}$, where $c(m_l)$ is the second component of the pair $(p_i, c(m_l))$ and represents the value of the token m_l at the place p_i .

For example, the 3-level inheritance tree for the knowledge base designed by the KRFPN (*Example 1*), where a token m_1 is initially at a place p_1 , is shown in Fig. 3.

Note that in order to simplify and make the notation shorter the nodes in the tree $\boldsymbol{\pi}_{pr}$, $p, r = 0, 1, 2, \dots$ contain only a component that is different to \emptyset (Fig. 3). This component contains information about the place where the token is and the token's value.

By using the components of the k -level inheritance tree, a node at the level $i - 1$, a labeled arc, a node at the level i (successor of the node at level $i - 1$), the functions α

and β , and a triplet named an *inheritance assertion* is formed. The strength of the assertion is defined as the value of the token at the successor node, i.e., as a product of the token value at the node at level $i - 1$ and the value of the association function of the corresponding transition.

The *inheritance paths*, starting from the root node of the inheritance tree and finishing at the leaves of the tree, represent sequences of the inheritance assertions. An inheritance path is interpreted as the conjunction of the inheritance assertions in which the redundant concepts connected by AND are omitted. The strength of an inheritance path is defined by the value of the token at the node that is a leaf of the inheritance tree.

In network-based knowledge representation schemes there is the well-known problem of the conflicting multiple inheritance [13], which in the KRFPN is expressed as follows: two inheritance paths are in conflict if the same concept inherits the mutually contradictory elements from D. Two inheritance paths are, also, in conflict if the same concept inherits the concept or property from D, but over contradictory relations from Σ . To resolve the situations involving conflicting multiple inheritance in the KRFPN we used Touretzky's principle of inferential distance ordering (PIDO) [13].

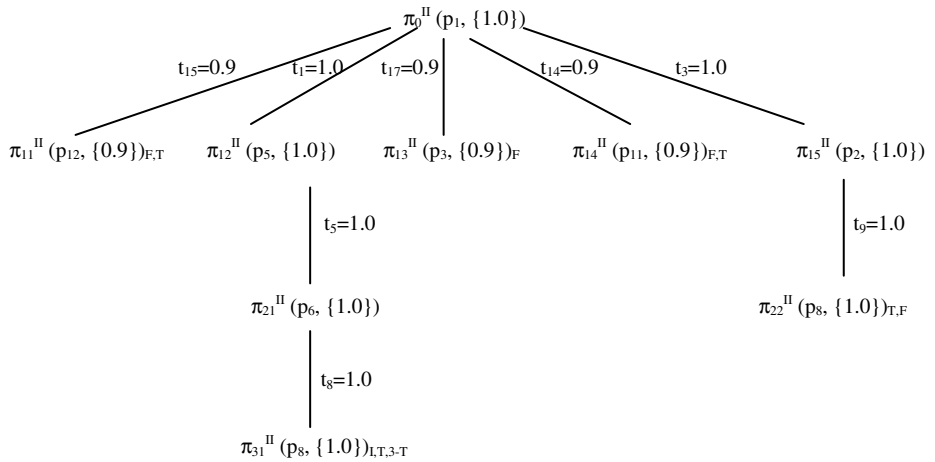


Fig. 3. The 3-level inheritance tree for the knowledge base (Example 1) (T denotes the terminal node, F denotes the frozen node and 3-T denotes the terminal node at level $k = 3$)

In situations when PIDO fails, procedure described in the *Step 9* below is applied. The intersection search algorithm for the KRFPN is presented as follows:

Input: Two concepts of interest, d_1 and d_2 , for which we want to determine possible relationships; the depth of the inheritance k ; $0 \leq k < \infty$, and $\lambda \in [0, 1]$.

Output: The relationships between the concepts d_1 and d_2 expressed by assertions (by means of the inheritance paths) from patriarch nodes to the intersection nodes.

In order to make the algorithm clearer, each of its steps will be illustrated for the following task: For the knowledge base (Fig. 2) find relationships between the concepts $d_1 = Scooby$ and $d_2 = Shaggy$. The depth of the inheritance is $k = 3$ and $\lambda = 0.1$.

Step 1. For the given concepts of interest, d_1 and d_2 , by using the inverse function α^{-1} , find the corresponding places p_i and p_j : $\alpha^{-1}: d_1 \rightarrow p_i$, $\alpha^{-1}: d_2 \rightarrow p_j$. If $d_1 \notin D$ or $d_2 \notin D$ stop the algorithm and send the message: “ d_u is an unknown concept, the relationships are unknown”; $u=1, 2$.

For our example: $\alpha^{-1}: Scooby \rightarrow p_3$, $\alpha^{-1}: Shaggy \rightarrow p_1$.

Step 2. Define the initial markings $\mu^I_0 = (\mu^I_1, \mu^I_2, \dots, \mu^I_n)$ and $\mu^{II}_0 = (\mu^{II}_1, \mu^{II}_2, \dots, \mu^{II}_n)$, where n is a cardinality of the set of places P :

$$\mu^I_k = \begin{cases} 1 & \text{for } k = i \\ 0 & \text{for all } k \neq i \end{cases} \quad \text{and} \quad \mu^{II}_k = \begin{cases} 1 & \text{for } k = j \\ 0 & \text{for all } k \neq j \end{cases}, \quad k = 1, 2, \dots, n$$

In our example: $\mu^I_0 = (0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0)$ and
 $\mu^{II}_0 = (1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)$.

Step 3. Define the initial distribution of tokens $\Omega^I_0 = \pi^I_0 = (\emptyset, \emptyset, \emptyset, \dots, \{(p_i, c(m^I_1))\}, \dots, \emptyset, \emptyset)$ and $\Omega^{II}_0 = \pi^{II}_0 = (\emptyset, \emptyset, \emptyset, \dots, \{(p_j, c(m^{II}_1))\}, \dots, \emptyset, \emptyset)$, and set $c(m^I_1) = c(m^{II}_1) = 1.0$;

$\Omega^I_0 = \pi^I_0 = (\emptyset, \emptyset, \{(p_3, c(m^I_1) = 1.0)\}, \dots, \emptyset, \emptyset)$ and
 $\Omega^{II}_0 = \pi^{II}_0 = (\{(p_1, c(m^{II}_1) = 1.0)\}, \emptyset, \emptyset, \dots, \emptyset, \emptyset)$

Step 4. For the initial distribution of tokens $\Omega^I_0 = \pi^I_0$ construct k levels of the inheritance tree $InhTree^I$. Fig. 4. depicts the $InhTree^I$.

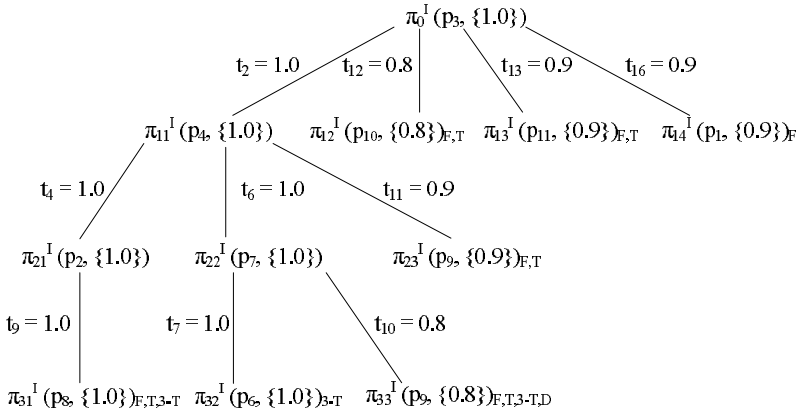


Fig. 4. The inheritance tree $InhTree^I$

Step 5. For the initial distribution of tokens $\Omega^{II}_0 = \pi^{II}_0$ construct k levels of the inheritance tree $InhTree^{II}$. The inheritance tree $InhTree^{II}$ is shown in Fig. 3.

Step 6. Find the nodes π^I_{pr} , $p, r = 0, 1, 2, \dots$ in $InhTree^I$ and π^{II}_{st} , $s, t = 0, 1, 2, \dots$ in $InhTree^{II}$ that match one another. Two nodes, one from $InhTree^I$ and another from $InhTree^{II}$, are the *matched nodes* if they have tokens in the same places, regardless of the number and the values of the tokens. These nodes are defined as the intersection

nodes. For example, the nodes $\pi_{pr}^I = (\emptyset, \emptyset, \emptyset, \{(p_4, c(m_1) = 0.8), (p_4, c(m_2) = 1.0)\}, \emptyset, \dots, \emptyset)$ and $\pi_{st}^II = (\emptyset, \emptyset, \emptyset, \{(p_4, c(m_1) = 0.6)\}, \emptyset, \dots, \emptyset)$ are matched nodes.

If there are no such nodes the algorithm stops and sends the message: “There are no relationships between the concepts (facts) d_1 and d_2 ”.

For our example (see Fig. 4 and Fig. 3):

$$\begin{aligned} (\pi_{0,}^I, \pi_{13,}^II): \pi_{0,}^I &= (p_3, \{1.0\}) \text{ and } \pi_{13,}^II = (p_3, \{0.9\}); \\ (\pi_{14,}^I, \pi_{0,}^II): \pi_{14,}^I &= (p_1, \{0.9\}) \text{ and } \pi_{0,}^II = (p_1, \{1.0\}); \\ (\pi_{13,}^I, \pi_{14,}^II): \pi_{13,}^I &= (p_{11}, \{0.9\}) \text{ and } \pi_{14,}^II = (p_{11}, \{0.9\}); \\ (\pi_{21,}^I, \pi_{15,}^II): \pi_{21,}^I &= (p_2, \{1.0\}) \text{ and } \pi_{15,}^II = (p_2, \{1.0\}); \\ (\pi_{31,}^I, \pi_{22,}^II): \pi_{31,}^I &= (p_8, \{1.0\}) \text{ and } \pi_{22,}^II = (p_8, \{1.0\}); \\ (\pi_{31,}^I, \pi_{31,}^II): \pi_{31,}^I &= (p_8, \{1.0\}) \text{ and } \pi_{31,}^II = (p_8, \{1.0\}); \\ (\pi_{32,}^I, \pi_{21,}^II): \pi_{32,}^I &= (p_6, \{1.0\}) \text{ and } \pi_{21,}^II = (p_6, \{1.0\}). \end{aligned}$$

Step 7. For all the matched nodes in $InhTree^I$ and $InhTree^{II}$ apply the semantic function α for the corresponding places to obtain a set of intersection concepts.

$\alpha: p_3 \rightarrow Scooby$, $\alpha: p_1 \rightarrow Shaggy$, $\alpha: p_{11} \rightarrow Hungry$, $\alpha: p_2 \rightarrow Cartoon_Character$, $\alpha: p_8 \rightarrow Live$, $\alpha: p_6 \rightarrow Mammal$;

The set of intersection concepts is: $\{Scooby, Shaggy, Hungry, Cartoon_Character, Live, Mammal\}$.

Step 8. Find the inheritance paths, starting from the root node (the patriarch node) of the inheritance tree $InhTree^I$ and finishing at the nodes (the leaves) of the k -level inheritance tree. Do the same for the inheritance tree $InhTree^{II}$.

The strength of an inheritance path is defined as the minimal value of the tokens in the leaf-node.

$InhTree^I$:

- (i) *Scooby is_a Cartoon_Dog AND is_a Cartoon_Character AND is_not_a Live*; strength = 1.0,
- (ii) *Scooby is_a Cartoon_Dog AND is_a Dog AND is_a Mammal*; strength = 1.0,
- (iii) *Scooby is_a Cartoon_Dog AND has_characteristic Talking*; strength = 0.9,
- (iv) *Scooby is_a Cartoon_Dog AND is_a Dog AND has_not_characteristic Talking*; strength = 0.8,
- (v) *Scooby has_color Brown*; strength = 0.8
Scooby has_characteristic Hungry; strength = 0.9,
- (vi) *Scooby is_behind_of Shaggy*; strength = 0.9.

$InhTree^{II}$:

- (i)' *Shaggy is_front_of Scooby*; strength = 0.9,
- (ii)' *Shaggy has_characteristic Hungry*; strength = 0.9,
- (iii)' *Shaggy is_a Cartoon_Character AND is_not_a Live*; strength = 1.0,
- (iv)' *Shaggy is_a Human AND is_a Mammal AND is_a Live*; strength = 1.0
- (v)' *Shaggy has_characteristic Wears_Clothes*; strength = 0.9

Step 9. If there are assertions, in one or both sets of the inheritance assertions, involving conflict due to multiple inheritance use the PIDO or, if that fails, make a decision on the basis of the more direct inheritance path. If two or more inheritance paths have the same length the concept inherits the property that corresponds to the stronger path. On the basis of the above criteria, remove the inheritance assertion that is the source of the conflict and all the inheritance assertions that follow it.

InhTree^I:

The inheritance paths (iii) and (iv) are in conflict: Can Scooby talk or not? Using the PIDO concept results in rejecting the inheritance path (iv) because the concept *Cartoon_Dog* is “nearer” to the concept *Scooby* than to the concept *Dog*.

InhTree^{II}:

The inheritance paths (iii)’ and (iv)’ are in conflict: Is Shaggy live or not? To resolve the conflict, the decision is made on the basis of the more direct inheritance path, because the PIDO concept failed (there is no hierarchical relationship between the concepts *Cartoon_Character* and *Human*). The more direct inheritance path is (iii)’.

Step 10. After the elimination of the conflicting assertions, the elements of the set of intersection concepts determined in Step 7 are identified in the inheritance paths. If a certain inheritance path does not contain any of the intersection concepts, the given path does not describe the relationship between the concepts of interest.

For our example:

- (i) *Scooby is_a Cartoon_Dog* AND *is_a **Cartoon_Character***; Always true,
- (ii) *Shaggy is_a **Cartoon_Character***; Always true,
- (iii) *Scooby is_a Cartoon_Dog* AND *is_a Cartoon_Character* AND *is_not_a **Live***; Always true,
- (iv) *Shaggy is_a Cartoon_Character* AND *is_not_a **Live***; Always true,
- (v) *Scooby is_a Cartoon_Dog* AND *is_a Dog* AND *is_a **Mammal***; Always true,
- (vi) *Shaggy is_a Human* AND *is_a **Mammal***; Always true,
- (vii) *Scooby has_characteristic **Hungry***; Very true,
- (viii) *Shaggy has_characteristic **Hungry***; Very true,
- (ix) *Scooby is_behind_of **Shaggy***; Very true,
- (x) ***Shaggy***; Always true,
- (xi) ***Scooby***; Always true,
- (xii) *Shaggy is_front_of **Scooby***; Very true.

(Note that the concepts corresponding to the intersection nodes are denoted bold).

4 Conclusion

An original intersection search procedure for the knowledge representation scheme based on the Fuzzy Petri Net theory, named KRFPN, is proposed. The procedure uses k -level inheritance trees that are generated on the basis of the dynamical properties of the scheme. The relationships between the concepts of interest, obtained by the proposed intersection search algorithm, are accompanied by the value of a linguistic variable expressing the degree of assurance for the relationship assigned to the transitions.

The very important properties of the proposed algorithm are that the k -level inheritance trees are finite and that the upper bound of the time complexity of the algorithm is $O(nm)$, where n is the number of the concepts (places) and m is the number of relations (transitions) in the knowledge base.

The program simulator for the KRFPN was developed and the fuzzy inference procedures, including the proposed intersection search algorithm, were tested on numerous examples [19].

References

1. Cardoso, J., Camargo, H. (eds.): Fuzziness in Petri Nets. Physica-Verlag, Heidelberg (1999)
2. Looney, C.G.: Fuzzy Petri Nets for Rule-based Decisionmaking. IEEE Trans. on the System, Man and Cybernetics 18(1), 178–183 (1988)
3. Chen, S.-M., Ke, J.-S., Chang, J.-F.: Knowledge Representation Using Fuzzy Petri Nets. IEEE Trans. on Knowledge and Data Engineering 2(3), 311–319 (1990)
4. Scarpelli, H., Gomide, F., Yager, R.R.: A Reasoning Algorithm for High Level Fuzzy Petri Nets. IEEE Trans. on Fuzzy Systems 4(3), 282–294 (1996)
5. Chen, S.-M.: Weighted Fuzzy Reasoning Using Weighted Fuzzy Petri Nets. IEEE Trans. on Knowledge and Data Engineering 14(2), 386–397 (2002)
6. Li, X., Lara-Rosano, F.: Adaptive Fuzzy Petri Nets for Dynamic Knowledge Representation and Inference. Expert Systems with Applications 19, 235–241 (2000)
7. Looney, C.G., Liang, L.R.: Inference via Fuzzy Belief Petri Nets. In: Proceed. of the 15th Int. Conf. on Tools with Artificial Intelligence (ICTAI 2003), pp. 510–514 (2003)
8. Lee, J., Liu, K.F.R., Chaing, W.: Model Uncertainty Reasoning With Possibilistic Petri Nets. IEEE Trans. on Systems, Man and Cybernetics–Part B: Cybernetics 33(2), 214–224 (2003)
9. Canales, J.C., Li, X., Yu, W.: Fuzzy Knowledge Learning via Adaptive Fuzzy Petri Net with Triangular Function Model, Intelligent Control and Automation. In: The Sixth World Congress on WCICA 2006, vol. 1, pp. 4249–4253 (2006)
10. Ha, M.-H., Li, J., Li, H.-J., Wang, P.: A New Form of Knowledge Representation and Reasoning. In: Proc. of the Fourth Int. Conf. on Machine Learning and Cybernetics, Guangzhou, pp. 2577–2582 (2005)
11. Guo-Yan, H.: Analysis of Artificial Intelligence Based Petri Net Approach to Intelligent Integration of Design. In: Proc. of the International Conference on Machine Learning and Cybernetics, pp. 1691–1695 (2006)
12. Shen, V.R.L.: Knowledge Representation Using High-Level Fuzzy Petri Nets. IEEE Trans. on Systems, Man, and Cybernetics–Part A 36(6), 1220–1227 (2006)
13. Touretzky, D.S.: The Mathematics of Inheritance Systems. Pitman, London (1986)
14. Quillian, M.R.: Word Concepts: A Theory and Simulation of Some Basic Semantic Capabilities. Behavioral Science 12(5), 410–430 (1967)
15. Shastri, L.: Semantic Networks: An Evidential Formalization and Its Connectionist Realization. Pitman, London (1988)
16. Peterson, J.L.: Petri Net Theory and Modeling of Systems. Prentice-Hall, Englewood Cliffs (1981)
17. <http://en.wikipedia.org/wiki/Scooby-Doo>
18. Ribarić, S., Pavešić, N.: Inference Procedures for Fuzzy Knowledge Representation Scheme. Applied Artificial Intelligence 23, 1, 16–43 (2009)
19. Zadrija, V.: Diploma Thesis, Faculty of EE and Computing. University of Zagreb (June 2008)

Parametric Uncertainty of Linear Discrete-Time Systems Described by Fuzzy Numbers

Petr Hušek*

Department of Control Engineering,
Faculty of Electrical Engineering,
Czech Technical University in Prague,
Technická 2, 166 27 Prague,
Czech Republic
husek@fel.cvut.cz
<http://dce.felk.cvut.cz>

Abstract. The paper deals with the problem of determination of stability margin of uncertain linear discrete-time systems with uncertainty described by fuzzy numbers. Nonsymmetric triangular membership functions describing the uncertainty of coefficients of characteristic polynomial are considered. The presented solution is based on transformation of the original problem to Hurwitz stability test and generalization of Tsytkin-Polyak plot.

1 Introduction

When dealing with the problems related to systems with parametric uncertainty classical robust analysis approach assumes that the uncertainty remains the same independently on the working conditions. It means that the worst case has to be considered and conservative results are obtained. However, in many practical situations the uncertainty varies, e.g. depending on operation conditions. In such a case the uncertainty interval is not fixed. One way how to characterize this dependency is a parameterization of uncertainty intervals by a confidence level. This parameter is usually tough to measure but it can be estimated by a human operator. If each parameter of a system is described in this way the system corresponds to a family of interval linear time-invariant systems parameterized by the confidence level.

Naturally, as in classical analysis of systems with structured uncertainty the parameterized uncertainty intervals can enter into the coefficients linearly, multilinearly, polynomially or even in more complicated manner. To handle such type of uncertain systems a mathematical framework is desired. Such a framework was proposed by Bondia and Picó in [1]. They adopted the concept of *fuzzy numbers* and *fuzzy functions*, see [2]. The approach interprets a set of intervals parameterized by a confidence level as a fuzzy number with its membership degree given

* This work has been supported by the Research Program MSM6840770038 and the project INGO 1P2007LA297 (sponsored by the Ministry of Education of the Czech Republic).

by this confidence level. It means that all the coefficients c_i are characterized by means of fuzzy numbers with membership functions $\alpha_i = \mu_{\tilde{c}_i}(c_i)$. When a confidence level α_i is specified then the coefficient interval is determined by the α_i -cut $[c_i]_{\alpha_i}$. If $\alpha_i = 1$ (the maximum confidence level – the system works in normal operating conditions) the coefficient c_i can take any value (crisp or interval) within the cores of \tilde{c}_i 's ($c_i = \ker\{\tilde{c}_i\}$). If $\alpha_i = 0$ (the minimum confidence level) the coefficient c_i is the interval equal to the support of \tilde{c}_i ($c_i \in \text{supp}\{\tilde{c}_i\}$). It is supposed that $\text{supp}\{\tilde{c}_i\}$ are finite sets, e.g. sigmoidal membership functions cannot be applied. Throughout the paper common confidence level α for all parameters will be used.

The question is what minimum confidence level α_{min} guarantees stability of the system under the assumption that the nominal system (i.e. for $\alpha = 1$) is Schur stable. A different definition for a measure of fuzzy system stability based on the degree of belief that a system is stable was proposed in [3]. In the sequel stable means Schur stable.

There exist two main approaches extensively used to solve the problem of robust stability analysis of systems with parametric uncertainty. The first one is called coefficient space concept or algebraic approach and it consists in algebraic computations with coefficients of characteristic polynomial and testing positivity of multivariate polynomial. The value set concept or frequency domain approach transforms multidimensional problem to testing two-dimensional sets in the complex plane. The latter is used in this paper.

2 Concept of Fuzzy Numbers

In this section we will formalize the concept of description of uncertain parameters by fuzzy numbers that conforms to the framework of interval polynomials.

Let us consider characteristic polynomial

$$\tilde{C}(z) = \tilde{c}_0 + \tilde{c}_1 z + \dots + \tilde{c}_n z^n \quad (1)$$

where the coefficients $\tilde{c}_k, k = 0, \dots, n$ are described by triangular membership functions (generally nonsymmetric). More precisely, considering common confidence level α , if triangular membership function with $\ker\{\tilde{c}_k\} = c_k^0, \text{supp}\{\tilde{c}_k\} = [c_k^-, c_k^+]$ characterizes the coefficient \tilde{c}_k then the functions

$$\begin{aligned} c_k^-(\alpha) &= (c_k^0 - c_k^-)\alpha + c_k^-, \\ c_k^+(\alpha) &= (c_k^0 - c_k^+)\alpha + c_k^+ \end{aligned} \quad (2)$$

determine the α -cut representation of polynomial (1) defined as an interval polynomial

$$\begin{aligned} \tilde{C}^\alpha(z) &= C(z, \alpha) = \sum_{k=0}^n c_k z^k, \\ c_k &\in [c_k^-(\alpha), c_k^+(\alpha)]. \end{aligned} \quad (3)$$

Let us suppose that the nominal (1-cut) polynomial $C(z, 1) = \sum_{i=0}^n c_i^0 z^i$ is stable. The task is to find stability margin of the polynomial (1), i.e. confidence level $\alpha_{\min} \in [0, 1]$ such that interval polynomial (3) is stable for $\alpha > \alpha_{\min}$ and unstable for $\alpha \leq \alpha_{\min}$.

The solution of this problem for continuous-time systems was stated in [4] with the help of Argoun stability test [5], which is graphical in nature, or in [6] using Kharitonov theorem and in [7] using the generalization of Tsyppkin-Polyak loci [8], [9].

When dealing with Schur stability of interval polynomials the solution is more complicated since Kharitonov theorem cannot be generally applied. The exception is the case when upper coefficients are fixed [10]. Some more counterparts on application of Kharitonov theorem to Schur stability are given in [11], [12] and [13]. Generally, a subset of exposed edges has to be tested for stability using Segment lemma [14] or using algebraic approach Jury test has to be performed [15]. Unfortunately, both approaches are not suitable for determination of stability margin. In this paper we propose different solution based on bilinear transformation of discrete-time polynomial to continuous-time one and using generalization of Tsyppkin-Polyak plot [8].

3 Transformation to Continuous-Time Polynomial

When dealing with stability determination of discrete-time systems with parametric uncertainty it is often more effective to perform stability analysis of equivalent continuous-time systems instead of the original discrete-time one. Using a bilinear transformation [16], [17] the determination of Schur stability of a given discrete-time polynomial can be converted to the determination of Hurwitz stability of an equivalent continuous-time polynomial. Utilizing this approach, the often more elaborated and powerful continuous-time design techniques can be applied to the discrete-time domain.

Lemma 1. *Schur stability margin of polynomial $\tilde{C}(z)$ is equal to Hurwitz stability margin of polynomial*

$$\tilde{D}(s) = (s - 1)^n \tilde{C}\left(\frac{s + 1}{s - 1}\right) = \tilde{d}_0 + \tilde{d}_1 s + \dots + \tilde{d}_n s^n. \quad (4)$$

Proof of the lemma 1 is obvious since the bilinear transformation $z = \frac{s+1}{s-1}$ maps the roots of $\tilde{C}(z)$ located inside, outside, or on the unit circle to the zeros of $\tilde{C}\left(\frac{s+1}{s-1}\right)$ located inside the open left half plane, inside the open right half plane or on the imaginary axis, respectively. It should be noted that the lemma 1 holds only if $\tilde{C}(z)$ has no roots at $z = -1$ (degree of $\tilde{D}(s)$ is equal to degree of $\tilde{C}(z)$). If $\tilde{C}(z)$ has some root at $z = -1$ a biquadratic transformation should be used [18].

The coefficients \tilde{d}_k , $k = 0, \dots, n$ of the polynomial (4) are linear affine functions of the coefficients \tilde{c}_k , $k = 0, \dots, n$ (1). It means that using bilinear transformation (4) we have transformed the task of determination minimum confidence

level guaranteeing Schur stability of interval polynomial to the task of determination minimum confidence level guaranteeing Hurwitz stability of interval polynomial with linear affine dependency on uncertain parameters.

4 Linear Affine Fuzzy Parametric Uncertainty

In the sequel we will consider polynomial

$$\tilde{D}(s) = \tilde{d}_0 + \tilde{d}_1 s + \cdots + \tilde{d}_n s^n \quad (5)$$

where the coefficients $\tilde{d}_i, i = 0, \dots, n$ are supposed to be linear affine functions of the parameters $\tilde{c}_k, k = 0, \dots, n$, i.e.

$$\tilde{d}_i = \beta_i + \sum_{k=0}^n \gamma_{ik} \tilde{c}_k, \quad \beta_i, \gamma_{ik} \in \mathfrak{R}. \quad (6)$$

The parameters $\tilde{c}_k, k = 0, \dots, n$ are described by nonsymmetric triangular membership functions sharing common confidence level α . If the triangular membership function with $\ker\{\tilde{c}_k\} = c_k^0, \text{supp}\{\tilde{c}_k\} = [c_k^-, c_k^+]$ describes the coefficient \tilde{c}_k then the linear functions

$$\begin{aligned} c_k^-(\alpha) &= (c_k^0 - c_k^-)\alpha + c_k^-, \\ c_k^+(\alpha) &= (c_k^0 - c_k^+)\alpha + c_k^+ \end{aligned} \quad (7)$$

characterize the linear interval polynomial

$$D(s, \alpha) = d_0(\alpha) + d_1(\alpha)s + \cdots + d_n(\alpha)s^n \quad (8)$$

where

$$\begin{aligned} d_i(\alpha) &= \beta_i + \sum_{k=0}^n \gamma_{ik} c_k, \quad i = 0, \dots, n, \\ c_k &\in [c_k^-(\alpha), c_k^+(\alpha)]. \end{aligned} \quad (9)$$

Let us suppose that the nominal (1-cut) polynomial $D(s, 1) = \sum_{i=0}^n d_i^0 s^i, d_i^0 = \beta_i + \sum_{k=0}^n \gamma_{ik} c_k^0$ is stable. We are looking for confidence level $\alpha_{\min} \in [0, 1]$ such that linear interval polynomial (8) is stable for $\alpha > \alpha_{\min}$ and unstable for $\alpha \leq \alpha_{\min}$.

In order to solve the problem a generalization of the Tsytkin-Polyak plot [8] will be used.

5 Stability Margin Determination

5.1 Zero Exclusion Theorem

Let \mathcal{Q} be a connected region in the $(n + 1)$ -dimensional space. Let us consider family of polynomials

$$\delta(s, \mathcal{Q}) = p_0 + \cdots + p_n s^n, \mathbf{p} = [p_0, \dots, p_n], \mathbf{p} \in \mathcal{Q}. \quad (10)$$

To derive the main result of this paper well-known boundary crossing theorem will be used.

Theorem 1 (*Boundary crossing theorem*) [14]. *The family of polynomials $\delta(s, \mathcal{Q})$ (10) of invariant degree is stable if and only if*

- a) *there exists a stable polynomial $\delta(s, \mathbf{p}^*)$, $\mathbf{p}^* \in \mathcal{Q}$,*
- b) *$j\omega \notin \text{roots}\{\delta(s, \mathcal{Q})\} \forall \omega \in \mathfrak{R}$.*

This intuitive result simply states the fact that the first encounter of polynomial with fixed degree (i.e. coefficient p_n does not include zero) with instability has to be on the boundary of stability domain. Computationally more efficient version of the boundary crossing theorem is formulated by the zero exclusion principle.

Theorem 2 (*Zero exclusion principle*) [14]. *The family of polynomials $\delta(s, \mathcal{Q})$ (10) of invariant degree is stable if and only if*

- a) *there exists a stable polynomial $\delta(s, \mathbf{p}^*)$, $\mathbf{p}^* \in \mathcal{Q}$,*
- b) *$0 \notin \delta(j\omega, \mathcal{Q}) \forall \omega \in \mathfrak{R}$.*

The set $\delta(j\omega, \mathcal{Q})$, $\omega \in \mathfrak{R}$ is called the value set. Due to symmetry of value sets it suffices to check zero exclusion for $\omega \geq 0$ only.

5.2 Main Result

Let us consider the polytope of polynomials of constant degree

$$Q(s, \rho) = A(s) + \rho \sum_{k=0}^n r_k B_k(s), r_k^- \leq r_k \leq r_k^+ \quad (11)$$

where

$$\begin{aligned} A(s) &= d_0^0 + d_1^0 s + \dots + d_n^0 s^n, d_i^0 = \beta_i + \sum_{k=0}^n \gamma_{ik} c_k^0, \\ i &= 0, \dots, n, \\ B_k(s) &= \gamma_{0k} + \gamma_{1k} s + \dots + \gamma_{nk} s^n, \\ r_k^- &= c_k^- - c_k^0, r_k^+ = c_k^+ - c_k^0, \\ k &= 0, \dots, n, \\ \rho &> 0. \end{aligned}$$

The family of polynomials (11) is usually written as

$$Q(s, \rho) = A(s) + \rho \sum_{k=0}^n [r_k^-, r_k^+] B_k(s). \quad (12)$$

Theorem 3 *The minimum confidence level preserving stability of (8)*

$$\alpha_{\min} = \max\{0, 1 - \rho_{\max}\} \quad (13)$$

where ρ_{\max} is maximum value of ρ preserving stability of (12) called stability margin.

Proof. Substituting $\alpha = 1 - \rho$ into (8) one obtains $D(s, 1 - \rho) = Q(s, \rho)$ from which (13) immediately follows.

Let us examine the value set of polynomial family (12) in some point $s = j\omega$,

$$Q(j\omega, \rho) = A(j\omega) + \rho \sum_{k=0}^n [r_k^-, r_k^+] B_k(j\omega). \tag{14}$$

Since $r_i, i = 0, \dots, n$ are interval parameters the value set is a parallelogram with $2(n + 1)$ vertices [19], in which there are $n + 1$ pairs of edges parallel with $B_k = B_k(j\omega), k = 0, \dots, n$, see Fig. 1. In particular, if the complex numbers A, B_0, \dots, B_n are defined as

$$\begin{aligned} A &= A(j\omega) = |A|e^{j\theta}, \\ B_k &= B_k(j\omega) = |B_k|e^{j\phi_k}, k = 0, \dots, n \end{aligned} \tag{15}$$

then the value set (14) equals to the set $A + \rho\mathcal{B}$ where

$$\mathcal{B} = \left\{ \sum_{k=0}^n r_k B_k : r_k^- \leq r_k \leq r_k^+ \right\}. \tag{16}$$

Due to zero exclusion theorem we need to examine when zero is excluded from value set $A + \rho\mathcal{B}$. The following result gives the answer.

Theorem 4 *The condition*

$$0 \notin A + \rho\mathcal{B}, \rho > 0 \tag{17}$$

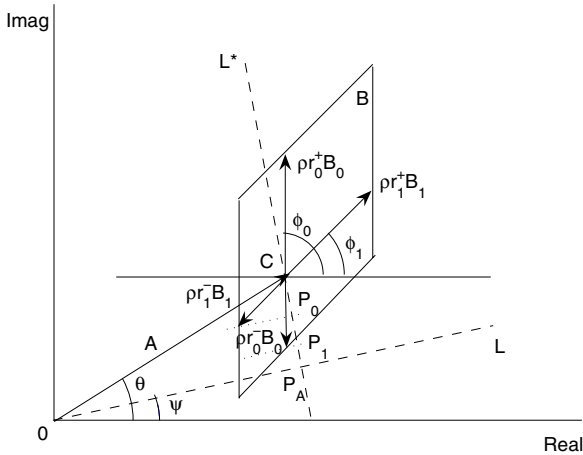


Fig. 1. Projection of the value set onto L^*

holds if and only if

$$\max_{0 \leq k \leq n} \frac{|A| |\sin(\theta - \phi_k)|}{\sum_{i=0}^n |r_i^* B_i| |\sin(\phi_i - \phi_k)|} > \rho, \\ \text{if } \sin(\phi_i - \phi_k) \neq 0 \text{ for some } i, k \quad (18)$$

where

$$r_i^* = \begin{cases} r_i^-, & \text{if } \text{sign}(\sin(\phi_i - \phi_k)) = \text{sign}(\sin(\phi_k - \theta)) \\ r_i^+, & \text{if } \text{sign}(\sin(\phi_i - \phi_k)) \neq \text{sign}(\sin(\phi_k - \theta)) \end{cases}$$

and

$$\max_{0 \leq k \leq n} \frac{|A|}{\sum_{i=1}^n |r_i^* B_i|} > \rho, \\ \text{if } \sin(\phi_i - \phi_k) = 0 \text{ and } \sin(\phi_k - \theta) = 0 \quad \forall i, k \quad (19)$$

where

$$r_i^* = \begin{cases} r_i^-, & \text{if } \theta = \phi_k, \\ r_i^+, & \text{if } \theta = -\phi_k. \end{cases}$$

Proof. Zero is excluded from the set $A + \rho\mathcal{B}$ if and only if there exists a line L which separates the set from the origin in complex plane. We will use the polygonal shape of value set and try to project the set into direction L^* which is orthogonal to the line L passing the origin at an angle ψ with the real axis, see Fig. [11](#). The length of the projection of the vector A into this direction is $|A| |\sin(\theta - \psi)|$ (the line CP_A in Fig. [11](#)). The total length of the projection of the set \mathcal{B} is

$$\rho \left((|r_0^+| + |r_0^-|) |\sin(\phi_0 - \psi)| + \dots \right. \\ \left. + (|r_n^+| + |r_n^-|) |\sin(\phi_n - \psi)| \right).$$

The line L separates the set $A + \rho\mathcal{B}$ from the origin if and only if the projection of A is greater than the part of the projection of \mathcal{B} (of the projections of each B_i) whose direction is opposite to the direction of the projection of A . In Fig. [11](#) these directions are $-B_0$ and $-B_1$ and the corresponding projections are the lines CP_0 and CP_1 . The total length of this part is

$$\rho (|r_0^*| |\sin(\phi_0 - \psi)| + \dots + |r_n^*| |\sin(\phi_n - \psi)|) \quad (20)$$

where

$$r_i^* = \begin{cases} r_i^-, & \text{if } \text{sign}(\sin(\phi_i - \psi)) = \text{sign}(\sin(\phi_k - \theta)) \\ r_i^+, & \text{if } \text{sign}(\sin(\phi_i - \psi)) \neq \text{sign}(\sin(\phi_k - \theta)) \end{cases}.$$

It means that if and only if there exists an angle $\psi \in [0, 2\pi)$ such that

$$|A| |\sin(\theta - \psi)| > \rho (|r_0^*| |\sin(\phi_0 - \psi)| + \dots \\ + |r_n^*| |\sin(\phi_n - \psi)|) \quad (21)$$

with r_i^* defined in [\(20\)](#) then the value set $A + \rho\mathcal{B}$ does not contain the origin. Because of polygonal shape of \mathcal{B} it suffices to test if the inequality [\(21\)](#) holds only for $\psi = \phi_i, i = 0, \dots, n$, which corresponds to the formula [\(17\)](#). The formula [\(19\)](#) solves the case when the value set \mathcal{B} degenerates to a line.

Since Theorem 4 holds for a polynomial of constant degree we have to investigate value of ρ_n corresponding to degree drop which for continuous-time polynomials causes loss of stability. By inspection of polytope (11) we obtain

$$\rho_n = \frac{d_n^0}{\sum_{k=0}^n r_k^* \gamma_{nk}} \quad (22)$$

where

$$r_k^* = \begin{cases} r_k^-, & \text{if } \text{sign}(d_n^0) = \text{sign}(\gamma_{nk}), \\ r_k^+, & \text{if } \text{sign}(d_n^0) \neq \text{sign}(\gamma_{nk}), k = 0, \dots, n. \end{cases}$$

In order to determine the stability margin ρ_{\max} of the polytope (12) we will look for maximum $\rho = \rho(\omega)$ for each $\omega \geq 0$ such that the inequalities (17) and (19) are satisfied. Then, if $\rho_{\min} := \inf_{\omega} \rho(\omega)$, stability margin of polynomial (12) (not necessarily of constant degree)

$$\rho_{\max} = \min\{\rho_{\min}, \rho_n\}. \quad (23)$$

The minimum confidence level α_{\min} preserving stability of (11) is then determined by (13).

6 Example

Let us consider a 6-th order discrete-time polynomial

$$\tilde{C}(z) = \tilde{c}_0 + \tilde{c}_1 z + \tilde{c}_2 z^2 + \tilde{c}_3 z^3 + \tilde{c}_4 z^4 + \tilde{c}_5 z^5 + \tilde{c}_6 z^6 \quad (24)$$

with the coefficients \tilde{c}_i , $i = 0, \dots, 6$ characterized by nonsymmetric triangular membership functions with

$$\begin{aligned} \ker\{\tilde{c}_0\} &= 0.0021, \quad \text{supp}\{\tilde{c}_0\} = [-0.0260 \ 0.0240] \\ \ker\{\tilde{c}_1\} &= 0.0350, \quad \text{supp}\{\tilde{c}_1\} = [-0.0110 \ 0.2050] \\ \ker\{\tilde{c}_2\} &= 0.2493, \quad \text{supp}\{\tilde{c}_2\} = [0.1940 \ 0.3440] \\ \ker\{\tilde{c}_3\} &= 0.9362, \quad \text{supp}\{\tilde{c}_3\} = [0.8852 \ 1.0192] \\ \ker\{\tilde{c}_4\} &= 1.9642, \quad \text{supp}\{\tilde{c}_4\} = [1.8542 \ 1.9932] \\ \ker\{\tilde{c}_5\} &= 2.1810, \quad \text{supp}\{\tilde{c}_5\} = [2.1100 \ 2.3300] \\ \ker\{\tilde{c}_6\} &= 1.0000, \quad \text{supp}\{\tilde{c}_6\} = [1.0000 \ 1.0000]. \end{aligned}$$

Let us determine the minimum confidence level α_{\min} preserving stability of polynomial (24).

Firstly we verify that the nominal polynomial

$$\begin{aligned} C(z, 1) &= \sum_{i=0}^6 c_i^0 z^i = z^6 + 2.1810z^5 + 1.9642z^4 \\ &\quad + 0.9362z^3 + 0.2493z^2 + 0.0350z + 0.0021 \end{aligned}$$

is Schur stable.

Now we will use bilinear transformation (4) to transform discrete-time polynomial (24) into equivalent continuous-time polynomial $\tilde{D}(s)$ (5). Using the α -cut representation (8) and notation (11), the nominal polynomial $D(s, 1)$ yields

$$D(s, 1) = A(s) = 6.3665s^6 + 17.9981s^5 + 21.0838s^4 \\ + 13.0984s^3 + 4.5509s^2 + 0.8383s + 0.0640.$$

The polynomials $B_k(s), k = 0, \dots, 6$ are given as

$$B_k(s) = (s + 1)^k (s - 1)^{6-k}$$

and $r_k, k = 0, \dots, 6$ as

$$\begin{aligned} r_0^- &= c_0^- - c_0^0 = -0.0281, & r_0^+ &= c_0^+ - c_0^0 = 0.0219 \\ r_1^- &= c_1^- - c_1^0 = -0.0460, & r_1^+ &= c_1^+ - c_1^0 = 0.1700 \\ r_2^- &= c_2^- - c_2^0 = -0.0553, & r_2^+ &= c_2^+ - c_2^0 = 0.0947 \\ r_3^- &= c_3^- - c_3^0 = -0.0510, & r_3^+ &= c_3^+ - c_3^0 = 0.0830 \\ r_4^- &= c_4^- - c_4^0 = -0.1100, & r_4^+ &= c_4^+ - c_4^0 = 0.0290 \\ r_5^- &= c_5^- - c_5^0 = -0.0710, & r_5^+ &= c_5^+ - c_5^0 = 0.1490 \\ r_6^- &= c_6^- - c_6^0 = -0.0000, & r_6^+ &= c_6^+ - c_6^0 = 0.0000. \end{aligned}$$

We are now looking for maximum ρ preserving stability of polytope (11). The corresponding frequency plot of $\rho(\omega)$ is shown in Fig. 2. From this plot we will find $\rho_{\min} = \inf_{0 < \omega < \infty} \rho(\omega) = 0.4575$. For $\omega = 0$ the value set degenerates to a line, using (19) we obtain $\rho_0 = \rho(0) = 0.1065$. The degree drop of polynomial (24) according to (22) occurs for $\rho_n = 13.4740$. Using (23) the stability margin ρ_{\max} is

$$\rho_{\max} = \min\{\rho_0, \rho_n, \rho_{\min}\} = 0.1065$$

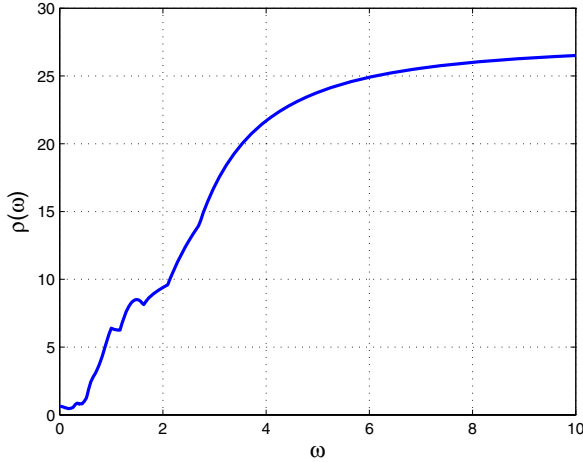


Fig. 2. Frequency plot of $\rho(\omega)$ of equivalent continuous-time polynomial to (24)

and the minimum confidence level α_{\min} preserving stability of (24),

$$\alpha_{\min} = \max\{0, 1 - \rho_{\max}\} = 0.8935.$$

7 Conclusion

In this paper an algorithm for determining stability margin of discrete-time linear systems with fuzzy parametric uncertainty is presented. The coefficients of characteristic polynomial are supposed to be characterized by nonsymmetric triangular membership functions. The paper extends the known results on continuous-time polynomials. The algorithm is based on bilinear transformation to continuous-time polynomial and generalization of Tsytkin-Polyak plot. The obtained result is demonstrated on an illustrative example.

References

1. Bondia, J., Picó, J.: Analysis of Systems with Variable Parametric Uncertainty Using Fuzzy Functions. In: Proc. of European Control Conference, ECC 1999 (1999)
2. Dubois, D., Prade, H.: Fuzzy Sets and Systems: Theory and Applications. Academic Press, Inc., London (1980)
3. Nguyen, H.T., Kreinovich, V.: How stable is a fuzzy linear system. In: Proc. 3rd IEEE Conf. on Fuzzy Systems, pp. 1023–1027 (1994)
4. Bondia, J., Picó, J.: Analysis of linear systems with fuzzy parametric uncertainty. Fuzzy Sets and Systems 135, 81–121 (2003)
5. Argoun, M.B.: Frequency domain conditions for the stability of perturbed polynomials. IEEE Trans. Automat. Control 10, 913–916 (1987)
6. Lan, L.: Robust stability of fuzzy-parameter systems. Automation and Remote Control 66, 596–605 (2005)
7. Hušek, P.: Generalized Tsytkin-Polyak locus. In: Proc. of 2nd IFAC Symposium on System, Structure and Control, SSSC 2004, pp. 78–83 (2004)
8. Tsytkin, Y.Z., Polyak, B.T.: Frequency domain criterion for robust stability of polytope of polynomials. In: Mansour, M., Balemi, S., Truol, W. (eds.) Control of Uncertain Dynamic Systems, pp. 113–124. Birkhauser, Basel (1992)
9. Mansour, M.: On robust stability of linear systems. Systems & Control Letters 22, 137–143 (1994)
10. Cieslik, J.: On possibilities of the extension of Kharitonov's stability test for interval polynomials to discrete-time case. IEEE Transactions on Automatic Control 32, 237–238 (1987)
11. Mansour, M., Kraus, F.J., Anderson, B.D.O.: Strong Kharitonov theorem for discrete systems. In: Milanese, M., Temp, R., Vicino, A. (eds.) Robustness in Identification and Control, vol. 22, pp. 113–124. Plenum Press, New York (1989)
12. Kraus, F.J., Anderson, B.D.O., Mansour, M.: Robust Schur polynomial stability and Kharitonov theorem. International Journal of Control 47, 1213–1225 (1988)
13. Greiner, R.: Necessary conditions for Schur-stability of interval polynomials. IEEE Transactions on Automatic Control 49, 740–744 (2004)
14. Bhattacharyya, S.P., Chapellat, H., Keel, L.: Robust Control: The Parametric Approach. Prentice-Hall, Inc., Upper Saddle River (1995)

15. Jury, E.: A modified stability table for linear discrete systems. *Proc. IEEE* 53, 184–185 (1965)
16. Ogata, K.: *Discrete-time control systems* (2nd ed.), 2nd edn. Prentice-Hall, Inc., Upper Saddle River (1995)
17. Shiomi, K., Otsuka, N., Inaba, H., Ishii, R.: The property of bilinear transformation matrix and Schur stability for a linear combination of polynomials. *Journal of Franklin Institute* 336, 533–541 (1999)
18. Jalili-Kharaajoo, M., Araabi, B.N.: The Schur stability via the Hurwitz stability analysis using a biquadratic transformation. *Automatica* 41, 173–176 (2005)
19. Kiselev, O., Lan, L., Polyak, B.: Frequency responses under parametric uncertainties. *Automation and Remote Control* 58, 645–661 (1997)

A Flexible Neuro-Fuzzy Autoregressive Technique for Non-linear Time Series Forecasting*

Alejandro Veloz^{1,2}, Héctor Allende-Cid², Héctor Allende², Claudio Moraga^{3,4},
and Rodrigo Salas¹

¹ Universidad de Valparaíso, Depto. de Ingeniería Biomédica, Valparaíso-Chile

² Universidad Técnica Federico Santa María, Depto. de Informática, Valparaíso-Chile

³ European Centre for Soft Computing, E-33600, Mieres-Spain

⁴ Dortmund University of Technology, 44221, Dortmund-Germany

Abstract. The aim of this paper is to simultaneously identify and estimate a non-linear autoregressive time series using a flexible neuro-fuzzy model. We provide a self organization and incremental mechanism to the adaptation process of the neuro-fuzzy model. The self organization mechanism searches for a suitable set of premises and consequents to enhance the time series estimation performance, while the incremental method selects influential lags in the model description.

Experimental results indicate that our proposal reliably identifies appropriate lags for non-linear time series. Our proposal is illustrated by simulations on both synthetic and real data.

Keywords: Non-linear Autoregressive Time Series, Neuro-fuzzy models, Flexible and Incremental learning.

1 Introduction

A common feature of all time series is that the current value is dependent of the previous ones, and a dynamic model is required for a proper description of the real system that generates the observations. In this way, an important problem in forecasting procedures is the empirical model building. This implies the need to select a proper set of lags operators on previous values of a time series to conform the independent variables that should be considered for an improved system modeling.

In system modeling and identification setting is assumed a dependence of current values of variables respect to the previous ones. But we have an extensive

* This work was supported in part by the Fondecyt 1070220 and DGIP-UTFSM research grants. The work of C. Moraga was partially supported by the Foundation for the Advancement of Soft Computing, Mieres, Asturias, Spain. E-mail addresses: avelozb@inf.utfsm.cl (A. Veloz), vector@inf.utfsm.cl (H. Allende-Cid), hallende@inf.utfsm.cl (H. Allende), mail@claudio-moraga.eu (C. Moraga) and rodrigo.salas@uv.cl (R. Salas)

list of possible forecasts, and we are not quite sure how many of these variables, and especially which ones, should be included at the final model. In general, the lags operators that conform the best set of explanatory variables are chosen with or without having to look at each possible subset.

On the other hand, Takagi-Sugeno-type neuro-fuzzy systems play an important role in many modeling and identification applications that exhibit complex relationships between variables related to a phenomenon of interest [9], and has been employed on dynamic autoregressive system identification (see [4], [6] and [8] for more details). In particular, this work is focussed on the development of a Takagi-Sugeno neuro-fuzzy model with flexible architecture for the conformation of an autoregressive framework that selects sets of the most explicative previous values from time series to improve forecasting performance.

The structure of this paper is organized as follows: in the next section, we briefly discuss the fundamentals of Non-linear Time Series Forecasting. In section 3 we present our proposed model. We give some experimental results to show the performance of the proposal in section 4. Finally, in the last section, we discuss the concluding remarks and we delineate some future works.

2 Non-linear Time Series Forecasting

The statistical approach to forecasting involves the construction of stochastic models to predict the value of an observation x_t using previous observations. This is often accomplished using linear stochastic difference equation models. By far, the most important class of such models is the linear autoregressive integrate moving average (ARIMA) model (see [2] for details). In [3] an extension of ARIMA to the non-linear case is presented.

An important class of Non-linear Time Series models is that of non-linear Autoregressive models (NAR) which is a generalization of the linear autoregressive (AR) model to the non-linear case. A NAR model obeys the equation $x_t = h(x_{t-1}, x_{t-2}, \dots, x_{t-p}) + \varepsilon_t$, where h is an unknown smooth non-linear function and ε_t is white noise, and it is assumed that $E[\varepsilon_t | x_{t-1}, x_{t-2}, \dots] = 0$. In this case the conditional mean predictor based on the infinite past observation is $\hat{x}_t = E[h(x_{t-1}, x_{t-2}, \dots) | x_{t-1}, x_{t-2}, \dots]$, with the following initial conditions $\hat{x}_0 = \hat{x}_{-1} = \dots = 0$.

3 Flexible Neuro-Fuzzy Autoregressive Technique (NFAR)

In this section we describe our proposed technique NFAR of a flexible neuro-fuzzy model developed to identify and estimate the non-linear autoregressive time series. The NFAR is an extension of ANFIS [5] and SONFIS [1]. During the learning procedure, our proposed model self-organizes its architecture in order to identify the number of fuzzy rules as well as the set of lags needed to describe and forecast the time series data.

Given the time series $\chi = \{x_t\}_{t=1}^T$, we assume that there exists an unknown regression function $\varphi(x_{t-1}, \dots, x_{t-p}) = \mathbb{E}[X_t | X_{t-1}, \dots, X_{t-p}]$ such that for any fixed value of x_{t-1}, \dots, x_{t-p} , the stochastic process is determined by $x_t = E[X_t | X_{t-1}, \dots, X_{t-p}] + \varepsilon_t$, where ε_t is a random variable with zero expectation and variance σ_ε^2 . The task of our neuro-fuzzy technique is to construct an estimator of the unknown function $\varphi(x_{t-1}, \dots, x_{t-p})$ by a set of K fuzzy *if-then* rules of Takagi and Sugeno type [9], i.e. for each k -th rule we have:

$$\text{Rule } k: \text{ If } x_{t-1} \text{ is } A_1^{(k)} \text{ and } \dots \text{ and } x_{t-p} \text{ is } A_p^{(k)}, \text{ then} \\ f_k(x_{t-1}, \dots, x_{t-p}, \Theta^{(k)}) = \theta_0^{(k)} + \theta_1^{(k)} x_{t-1} + \dots + \theta_p^{(k)} x_{t-p} = [\Theta^{(k)}]' \tilde{\mathbf{x}}$$

where the input features are given by the augmented vector of time-lag features $\tilde{\mathbf{x}} = [1, x_{t-1}, \dots, x_{t-p}]'$ and the vector of parameters of the k -th rule is $\Theta^{(k)} = [\theta_0^{(k)}, \theta_1^{(k)}, \dots, \theta_p^{(k)}]'$. The operator v' denotes the transpose of the vector v . These rules are modeled with a neuro-fuzzy system.

3.1 Architecture of the NFAR Model

The base architecture of NFAR was inspired on the structure of ANFIS proposed by Jyh-Shing Roger Jang in 1993 [5]. It is a fuzzy inference system implemented in the adaptive framework of neural networks and allows us to construct an input-output mapping based on both human intelligence and data samples.

The *fuzzification layer* computes the degree to which a given input x_i satisfies the linguistic label $A_i^{(k)}$. The output of the node is given by the membership function $\mu_{A_i^{(k)}}(x_i)$. In this work we use the gaussian-type membership function, $\mu_{A_i^{(k)}}(x_i; \eta_i^{(k)}) = \exp \left[- \left((x_i - \nu_i^{(k)}) / \sigma_i^{(k)} \right)^2 \right]$, where $\eta_i^{(k)} = \{\nu_i^{(k)}, \sigma_i^{(k)}\}$ are the *premise parameters* of the linguistic label $A_i^{(k)}$. The *generalized "AND layer"* consists of T - *norm* operators that perform the generalized AND. Each node of this layer represents the firing strength of some specific rule. We use the product as a T - *norm*, i.e., $w_k = w_k(\mathbf{x}; \eta^{(k)}) = \mu_{A_1^{(k)}}(x_1) \times \dots \times \mu_{A_d^{(k)}}(x_d)$, $k = 1..K$. The *normalization layer* computes the *normalizing firing strengths* of the weights of the previous layer as $\bar{w}_k = \bar{w}_k(\mathbf{x}; \eta^{(k)}) = w_k / \sum_{j=1}^K w_j$, $k = 1..K$. The *consequent layer* computes the weighted hyperplane that approximates the non-linear mapping, i.e., $\bar{f}_k(\mathbf{x}; \eta^{(k)}, \Theta^{(k)}) = \bar{w}_k(\mathbf{x}; \eta^{(k)}) f_k(\mathbf{x}; \Theta^{(k)}) = \bar{w}_k [\Theta^{(k)}]' \tilde{\mathbf{x}}$, where $\Theta^{(k)}$ is the consequent vector of parameters. Finally, the *network output* consists in a single node that computes the overall output as the summation of all the incoming signals:

$$g(\mathbf{x}; \eta, \Theta) = \sum_{k=1}^K \bar{w}_k(\mathbf{x}; \eta^{(k)}) f_k(\mathbf{x}; \Theta^{(k)}) \quad (1)$$

where $\eta = (\eta_1^{(1)}, \dots, \eta_d^{(K)})'$ and $\Theta = (\Theta^{(1)}, \dots, \Theta^{(K)})'$ correspond to the premise and consequent set of parameters respectively. To estimate these parameters,

the classical ANFIS employs a hybrid learning procedure that uses backpropagation learning algorithm to determine premise parameters η and the Least Mean Square (LMS) estimation procedure to determine the consequent parameters Θ . Please refer to reference [5] for further details.

3.2 Self-organization Process

During the learning procedure, our proposed model self organizes its architecture in order to automatically identify the number of rules needed to model the available data. The self-organization learning procedure consists in two stages. In the first stage we construct a base model with a predefined number of rules (nodes) and we iteratively estimate the premise η and consequent Θ parameters. The aim of the second stage is to select the most suitable number of membership functions by means of a self organization process. To accomplish this task the method applies three types of operators: *Grow Net*, *Split Membership Functions* and *Vanish Membership Functions*. Before applying any operator, the current base model is frozen meaning that none of its parameters can be any longer updated during this stage.

The *Grow Net* operator consists in adding new rules, i.e., a new membership function is incorporated to the premises of each input feature and together with a new hyperplane to the consequent. For each input \mathbf{x} of the training data we compute the firing strength w_k for all the K rules. If the maximum of these strengths is less or equal than a user-defined threshold δ to the power of d , where d is the dimension of the input space, i.e., $\max_{k=1..K} w_k \leq \delta^d$, then we say that the sample (\mathbf{x}, y) is not well-modeled by the current model. The sample is added to a “bad samples” set, together with label of its best matching rule, i.e.,

$$w_\kappa = \arg \max_{k=1..K} w_k(\mathbf{x}, \eta^{(k)}) \quad (2)$$

After having revised all the training data, we group the data into the set \mathcal{V}_κ according to their best matching rule w_κ . For each group that has more than N_{grow} samples, where N_{grow} is user-defined, we construct and add a new membership function for each dimension $\mu_{A_i^{(K+1)}}(x_i), i = 1..d$, with the premise parameters initialized with the mean, $\nu_i^{(K+1)}$, and standard deviation, $\sigma_i^{(K+1)}$, of the samples belonging to this group. The consequent parameters of the new rule are randomly initialized.

The *Split Membership Functions* operator consists in splitting a rule with bad performance into two new rules. To evaluate the rule performance, the training set is partitioned in K sets where the sample (\mathbf{x}, y) is assigned to the set \mathcal{V}_k if its best matching rule (2) is w_k . For each set we compute the mean square error $E_k = \frac{1}{N_k} \sum_{(\mathbf{x}, y) \in \mathcal{V}_k} (y - g(\mathbf{x}; \eta, \Theta))^2$, where N_k is the number of samples belonging to the set \mathcal{V}_k . If the performance of the rule k , E_k , is higher than a user defined threshold ϵ and N_k is higher than the minimum required samples N_{split} , where N_{split} is user-defined, then the rule is divided into two new rules. If the premise parameters of the k -th rule are ν and σ , then the premise parameters of

the new rules are $\nu_i^{(K+1)} = \nu - \frac{\sigma}{2}$, $\sigma_i^{(K+1)} = \frac{\sigma}{2}$, $\nu_i^{(K+2)} = \nu + \frac{\sigma}{2}$; and $\sigma_i^{(K+2)} = \frac{\sigma}{2}$. The consequent parameters are initialized randomly. After the inclusion of the new rules, the k -th rule is eliminated.

The *Vanish Membership Functions* operator consists in eliminating rules that model less than N_{vanish} sample data, where N_{vanish} is user-defined. To accomplish this, we introduce an age_k variable that starts from zero and is increased by one if the rule models no data, i.e. if the set \mathcal{V}_k is empty, and the rule is unfrozen. If the age variable of the k -th rule reaches the threshold λ , then the k -th rule is eliminated, where λ is user-defined. If the set \mathcal{V}_k is no longer empty, then the age_k variable is set back to 0.

After all three operators are applied to the model, all the unfrozen parameters (rules) are updated. After this step, the whole net architecture is frozen. The operators and the training steps are applied iteratively until the learning algorithm self organizes and stabilizes satisfying some user's performance criterion. Finally, all the parameters (frozen and unfrozen) of the network are updated in the last iteration.

3.3 Incremental Lags Identification

For the proposed NFAR method, we will explore two variants depending on the search mechanism employed to identify the more significant lags of the time series. The first method, called *Sequential NFAR* (NFAR-S), has a sequential incremental identification algorithm, where the search is accomplished sequentially from the second lag $t - 2$ until the user defined last possible lag $t - p_{max}$. The second method, called *Informative NFAR* (NFAR-I), starts the search from the most significant lag between $t - 2$ and $t - p_{max}$, and the technique incrementally adds the new feature to the neuro-fuzzy structure.

Both incremental algorithms work as follows. We assume that lag $t - 1$ is the most relevant one to predict the expected value of X_t , and we construct an train a Neuro-Fuzzy model with the self organization process explained in section 3.2. In this way, we obtain the model $\hat{x}_t = g(x_{t-1}; \eta^{(1)}, \Theta^{(1)})$ as given in equation (II), where the number K of rules was found automatically with the self organization process. After the training process of the base model, the current structure and parameters are frozen.

Now, suppose that the feature x_{t-q} is going to be added to the NFAR model. The new regressor feature x_{t-q} is incorporated to the architecture, together with its membership functions $\mu_{A_q^{(k)}}(x_{t-q}; \eta_q^{(k)})$, $k = 1..K$, with linguistic labels $A_q^{(k)}$, $k = 1..K$. All the consequents hyperplanes are extended with the new input feature, i.e., the function of the k -th rule, $k = 1..K$, becomes $f_k(\mathbf{x}, x_{t-q}, \Theta^{(k)})$ where \mathbf{x} symbolizes the previously selected features, and the vector of parameters $\Theta^{(k)} = [\theta_0^{(k)}, \theta_1^{(k)}, \dots, \theta_q^{(k)}]'$ has one more element $\theta_q^{(k)}$ for each rule. The consequent parameters are initialized randomly while the premise parameters are initialized with the values of the parameters of the last added premise, i.e, $\eta_q^{(k)} = \eta_r^{(k)}$, $k = 1..K$, where x_{t-r} was the last feature incorporated to the structure. The unfrozen parameters are trained with the available data.

To decide if the feature x_{t-q} will be added to the net architecture, we compute the performance of the model with the mean square error. If the error is reduced by at least a user-defined $\gamma\%$, then the new feature is accepted otherwise we proceed the search with next feature for the NFAR-S case or we stop searching for the NFAR-I case.

Once the feature x_{t-q} was included, all the parameters (frozen and unfrozen) of the network are updated. After the last training process, the net architecture is frozen again.

Notice that NFAR-I method searches the most significant lag by testing all the possible remaining (not included) feature lags x_{t-q} , $q = 2..p_{max}$, at the same time. Meanwhile the NFAR-S searches sequentially from $q = 2$ to $q = p_{max}$.

4 Experimental Results and Discussion

In this section we study the performance of our proposed model with its two variants, the sequential NFAR-S and the informative NFAR-I, compared to the Adaptive-Network-Based Fuzzy Inference System (ANFIS) and the classical Feedforward Artificial Neural Network (FANN). The experiments were executed with both synthetic and real datasets.

For the synthetic experiment we have created two synthetic time series. The simulated time series consist in a linear Autoregressive (AR) process, $x_t = 0.5x_{t-1} - 0.6x_{t-2} - 0.1x_{t-7} + 0.2x_{t-10} + \varepsilon_t$, and a non-linear Autoregressive (NAR) process: $x_t = 0.6x_{t-1}e^{-4*x_{t-2}^2} - 0.3x_{t-12}e^{-8x_{t-13}^2} + \varepsilon_t$. In both cases, ε_t is a Gaussian noise with zero mean and variance $\sigma_\varepsilon^2 = 0.1$. 5000 data samples were drawn with initial conditions of $x_0 = x_1 = \dots x_p = 0$ and only the last 1000 samples were used for the experimentation to avoid initialization problems, from where 500 samples were used for training, 250 for validation and 250 for testing.

In the real experiment we tested the algorithms with three real time series: a) The International Airline Passengers (Series G) of Box-Jenkins-Reinsel time series data set [2], with 120 samples for training, 12 samples for validation and 12 samples for testing. b) The Balloon Time Series data sets obtained from the StatLib archive (see [7]), with 1501 samples for training, 250 samples for validation and 250 samples for testing. c) The Laser Time Series, with 901 samples for training, 100 samples for validation and 100 samples for testing.

The parameters chosen empirically for the NFAR model are: $\epsilon = 0.3$, $N_{split} = 30$, $\delta = 0.7$, $N_{grow} = 20$, $\lambda = 2$, $N_{vanish} = 20$, $\gamma = 10\%$ and $p_{max} = 15$. The lags selected for the ANFIS and FANN models for the synthetic experiments are the same as in the original structure, i.e., the lags $(t-1)(t-2)(t-12)(t-13)$ and $(t-1)(t-2)(t-7)(t-10)$ for the first and second synthetic time series respectively. On the other hand, the lags selected for the ANFIS and FANN models for the real experiments coincide with the lags found by the best NFAR model, i.e., the lags $(t-1)(t-12)(t-13)$, $(t-1)$ and $(t-1)(t-2)(t-5)$ for the Airline, Balloon and Laser real time series respectively. Furthermore, the number of rules selected for the ANFIS, and the number of hidden neurons of the FANN were selected according to the NFAR methods.

The synthetic and real data sets were separated in training, validation and test sets. For each data sets, the models were executed 10 times and we computed the average of the mean square error of the test set (*MSE-Test*), the mean number of rules (*Rules*), the minimum mean square error of the test set (*Min-Test*), and the best set of lags (*Lags*). The experimental results are shown in table 1.

In both, synthetic and real time series data sets, our proposed NFAR models, NFAR-S and NFAR-I, outperformed significantly the FANN and ANFIS models in every measurement. In the Airline and Laser data sets, the improvement obtained by the NFAR methods were statistically significant. Notice that for the rigid structures models (FANN and ANFIS), the architecture was previously determined according to the best structure obtained by the NFAR in the real case, and by the structure of the underlying time series generation process for the synthetic case.

As we can further notice, the results obtained by the NFAR-S and NFAR-I are very similar. However, in some cases the NFAR-I obtained lower test errors than the NFAR-S, mainly due to the search strategy of the optimal set of lags. The NFAR-I is advocated to incorporate the most relevant lags first, while the NFAR-S looks sequentially for all the lags that improves the performance.

Table 1. Summary results of the performance evaluation of the FANN, ANFIS, NFAR-S and NFAR-I with the synthetic and real time series data sets

Dataset	Method	Rules	Lags	MSE -Test	Min-Test
Linear Synthetic AR	FANN	3	(t-1),(t-2),(t-12),(t-13)	$29.1 \cdot 10^{-5} \pm 1.9 \cdot 10^{-6}$	$28.8 \cdot 10^{-5}$
	ANFIS	3	(t-1),(t-2),(t-12),(t-13)	$10.0 \cdot 10^{-5} \pm 1 \cdot 10^{-6}$	$11.0 \cdot 10^{-5}$
	NFAR-S	3	(t-1),(t-12)	$9.0 \cdot 10^{-5} \pm 0$	$9.0 \cdot 10^{-5}$
	NFAR-I	2	(t-1),(t-12)	$9.0 \cdot 10^{-5} \pm 0$	$9.0 \cdot 10^{-5}$
Non-linear Synthetic AR	FANN	3	(t-1),(t-2),(t-7),(t-10)	$2.2 \cdot 10^{-4} \pm 1.1 \cdot 10^{-6}$	$2.2 \cdot 10^{-4}$
	ANFIS	3	(t-1),(t-2),(t-7),(t-10)	$1.1 \cdot 10^{-4} \pm 2.3 \cdot 10^{-6}$	$1.1 \cdot 10^{-4}$
	NFAR-S	3	(t-1),(t-2)	$1.0 \cdot 10^{-4} \pm 0$	$1.0 \cdot 10^{-4}$
	NFAR-I	2	(t-1),(t-2)	$1.1 \cdot 10^{-4} \pm 0$	$1.1 \cdot 10^{-4}$
Airline	FANN	5	(t-1),(t-12),(t-13)	35047 \pm 723	33876
	ANFIS	5	(t-1),(t-12),(t-13)	232274 \pm 0	2322744
	NFAR-S	6	(t-1),(t-9),(t-10), (t-11),(t-12),(t-13)	513 \pm 7	507
	NFAR-I	5	(t-1),(t-12),(t-13)	493 \pm 15	468
Balloon	FANN	4	(t-1)	$468.6 \cdot 10^{-3} \pm 3.2 \cdot 10^{-3}$	$463.9 \cdot 10^{-3}$
	ANFIS	4	(t-1)	$8.1 \cdot 10^{-3} \pm 2.6 \cdot 10^{-4}$	$7.5 \cdot 10^{-3}$
	NFAR-S	5	(t-1)	$8.1 \cdot 10^{-3} \pm 1.9 \cdot 10^{-4}$	$7.8 \cdot 10^{-3}$
	NFAR-I	4	(t-1)	$8.0 \cdot 10^{-3} \pm 2.0 \cdot 10^{-4}$	$7.7 \cdot 10^{-3}$
Laser	FANN	18	(t-1),(t-2),(t-5)	8543 \pm 35	8508
	ANFIS	18	(t-1),(t-2),(t-5)	7383 \pm 47	7254
	NFAR-S	28	(t-1),(t-2),(t-5)	923 \pm 536	142
	NFAR-I	18	(t-1),(t-2),(t-5)	79 \pm 31	35

5 Conclusion

In this paper we were able to give a solution to the problem of time series modeling by a Takagi-Sugeno type neuro-fuzzy system with a flexible self-organizing structure. The model was able to effectively identify the best set of explanatory variables in an incremental fashion. During the learning procedure, the NFAR model self organizes its architecture in order to automatically identify the number of rules together with the set lags needed to model the available data.

In the experimental studies with synthetic and real time series data sets, the proposed model showed a superior prediction performance than the classical ANFIS and FANN models, both with “rigid” architectures. Moreover, the NFAR model has the advantage of automatically search the best set of rules together with the most significant lags. This is a useful improvement in applications where an unexperienced user does not know which lags to use in order to forecast time series, leaving this task to a fully automated method. Two search methods to identify the most explicative variables were explored: the sequential and the informative, where the latter showed similar or better performance results.

Future works are needed in order to determine the values of the set of parameters of the NFAR model, possible improvements can be accomplished with evolutive frameworks. Another problem that should be attacked is to generalize our technique to non-stationary environments where concept drifts can occur. Several other search methods for the optimal set of lags should be explored in order to improve the performance of the current model.

References

1. Allende-Cid, H., Veloz, A., Salas, R., Chabert, S., Allende, H.: Self-organizing neuro-fuzzy inference system. In: Ruiz-Shulcloper, J., Kropatsch, W.G. (eds.) CIARP 2008. LNCS, vol. 5197, pp. 429–436. Springer, Heidelberg (2008)
2. Box, G.E.P., Jenkins, G.M., Reinsel, G.C.: Time series analysis, forecasting and control, 3rd edn. Prentice-Hall, Englewood Cliffs (1994)
3. Chow, T.W.S., Leung, C.-T.: Nonlinear autoregressive integrated neural network model for short-term load forecasting. IEE Proc. Generation, Transmission and Distribution 143(5), 500–506 (1996)
4. Golob, M., Tovornik, B.: Input-output modelling with decomposed neuro-fuzzy ARX model. Neurocomputing 71, 875–884 (2008)
5. Jang, J.-S.R.: ANFIS: Adaptive-network-based fuzzy inference system. IEEE, Transaction on Systems, Man and Cybernetics 23(3), 665–685 (1993)
6. Luna, I., Soares, S., Ballini, R.: A constructive-fuzzy system modeling for time series forecasting. In: Proceedings of International Joint Conference on Neural Networks, pp. 2908–2913. IEEE, Los Alamitos (2007)
7. Department of Statistics at Carnegie Mellon University, StatLib - datasets archive, <http://lib.stat.cmu.edu/>
8. Serra, G., Bottura, C.: An IV-QR algorithm for neuro-fuzzy multivariable online identification. IEEE Transactions on Fuzzy Systems 15(2), 200–210 (2007)
9. Takagi, T., Sugeno, M.: Derivation of fuzzy control rules from human operator’s control actions. In: Proc. IFAC Symp. Fuzzy Information, Knowledge Representation and Decision Analysis, pp. 55–60 (1983)

Multiagent Security Evaluation Framework for Service Oriented Architecture Systems*

Grzegorz Kołaczek

Institute of Informatics
Wroclaw University of Technology, Wroclaw, Poland
Grzegorz.Kolaczek@pwr.wroc.pl

Abstract. As more and more organizations use the Service Oriented Architecture (SOA) to design and implement their information systems also the systems' architects need the more intelligent and reliable tools. The complexity, modularity and heterogeneity of the information systems make the security evaluation process difficult. The proposed method uses multiagent approach as the most promising direction of the research. As the security evaluation requires the precise definition of the set of evaluation criteria the basic criteria for each functional layer of SOA have been presented. Also, the paper presents two algorithms where the first can be used separately for each of the particular layer of SOA and the second serves for the calculation of the generalized SOA system security level.

1 Introduction

Most organizations deliver their business processes using information technology (IT) applications. Many different software tools are used to capture, transform or report business data. Their role may be for example to structure, define and transform data or to enable or simplify communication. Each such interaction with an IT asset can be defined a service. The set of delivered from the business processes services provide the incremental building blocks around which business flexibility revolves. In this context, Service Oriented Architecture (SOA) is the application framework that enables organizations to build, deploy and integrate these services independent of the technology systems on which they run.[8] In SOA, applications and infrastructure can be managed as a set of reusable assets and services. The main idea about this architecture was that businesses that use SOA can respond faster to market opportunities and get more value from their existing technology assets.[9]

The final success of the SOA concept can be obtained if many groups, both internal and external to the organization, contribute to the execution of a business process. Because in most cases the most valuable and also sensible part of each organization is information, a business partner is much more willing to share information and data assets, if it knows that these assets will be protected and their integrity maintained. Business partners will also be more likely to use a service or process from another group if it has assurance of that asset's integrity and security, as well as reliability and

* The research presented in this paper has been partially supported by the European Union within the European Regional Development Fund program no. POIG.01.03.01-00-008/08.

performance. Therefore ensuring security is a one of the most crucial elements while putting SOA approach into practice.

The paper is structured as follows. The second section presents the general motivation and related work to the problems of security level evaluation and service oriented architecture. Next section describes a few basic notions in a security governance and after that the some examples of basic requirements for security evaluation in SOA. The forth section brings the main contribution – the description of multiagent framework for security evaluation in SOA systems. The last section consists of the conclusion and the direction of future research.

2 Motivation and Related Work

A mobile agent is a composition of computer software and data which is able to move from one host to another autonomously and continue its execution on the destination host. Mobile agent technology can reduce the bandwidth requirement and tolerate the network faults - able to operate without an active connection between clients and server. As the security evaluation process must be accurate and efficient, these basic features relevant to agent and multiagent systems are the main motivation for many researchers to apply multiagent approach to the tasks related to system security. The second premise in this case is the correspondence of the multiagent environment to SOA systems. Multiagent systems are composed from the number of autonomous and mobile entities that are able to act both cooperatively and separately. The fundamental concept for SOA system is service – entity that could be evoked individually as well as in cooperation with other services. And at last, both multiagent and SOA systems tend to act in heterogenic and highly distributed environment.

As the number of SOA implementation grows the concerns about SOA systems security also increases. The literature related to the security of SOA focuses on problems with threat assessment, techniques and functions for authentication, encryption, or verification of services.[1],[2],[6] Some other works focus on high level modeling processes for engineering secure SOA [4],[9] with trust modeling [7], identity management and access control [12][10]. Many studies focus on secure software design practices for SOA, with special interest in architectural or engineering methodologies as the means to create secure services. [3],[5]

To the author best knowledge, the proposed in this paper framework is the first that introduces the multiagent approach to the SOA security level evaluation. The other important and novel issues addressed in this work are the personalization of the security level evaluation process, multilevel security evaluation, support for continuous and automate security evaluation.

3 SOA Security Governance

There are several standards and mechanisms that have been elaborated to provide and to maintain the high security level of SOA systems. The basic solutions address the problems of confidentiality and integrity of data processed by SOA system. Because of the network context of SOA and the multilevel security risk related to the layers of

ISO/OSI network model, there are several solutions that offer data protection mechanisms at the corresponding level to each network layer.

The most commonly used and described are standards and protocols from the application layer that are maintained by OASIS consortium. These solutions have been worked out to support the development of web services and so also SOA systems. The other type of the protection methods, mechanisms and protocols like for example IPv6 are common for all network applications and can be used in SOA systems as well in any other type of software.

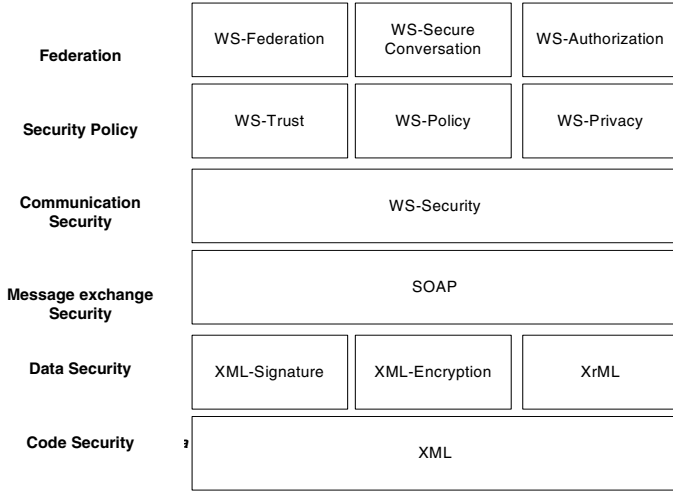


Fig. 1. SOA application layer related security protocols

3.1 Evaluation of SOA Security

The security evaluation process should be based on some formal prerequisites. This means that the security evaluation must be objective to guarantee the repeatability and universality of the evaluation results. So, there must be defined notion of the security

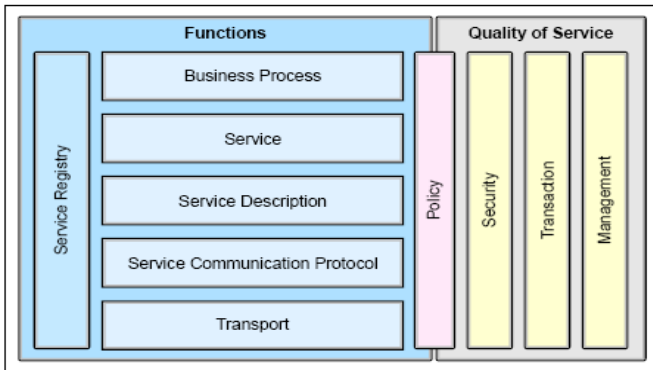


Fig. 2. The layered model of Service Oriented Architecture

measure. There are some confusion about this notion. The first problem is that the security measure does not have any specific unit. The other difficulties are: security level has no objective grounding but it only in some way reflects the degree in which our expectation about security agree with reality, security level evaluation is not fully empirical process, etc.

As the SOA system can be defied by its five functional layers (fig.2.) the correspondent definition of SOA security requirements for security evaluation process should address the specific security problems within each layer. Some elements from a set defining security requirements for the SOA layers has been presented in Table 1. The complete list can be found in [5].

Table 1. The functional and non-functional security evaluation requirements for each of the SOA functional layers (selection)

SOA Layer	Evaluate/verify/test
Policy and Business Processes	<ul style="list-style-type: none"> – Policy consistency – Policy completeness – Trust management – Identity management
Service	<ul style="list-style-type: none"> – Identification of the services – Authentication of the services – Management of security of the complex services
Service Description	<ul style="list-style-type: none"> – Description completeness – Availability – Protection from attacks
Service Communication Protocol	<ul style="list-style-type: none"> – Confidentiality – Authentication – Norms compliance
Transport	<ul style="list-style-type: none"> – Availability – Protection from attacks – Integrity

4 Multiagent Framework for SOA Security Evaluation

There are several different problems considering the SOA security level evaluation process. The most crucial, as it has been stated in the earlier sections are: the complexity of the architecture, multilevel relationships between the system components and heterogenic environment. Each security level evaluation method and tool must take into account all these factors and apply appropriate solutions for them to provide the accurate final results of the security evaluation process.

4.1 The Architecture of Multiagent System for SOA Security Level Evaluation

This section presents the main assumptions about SOA systems security evaluation framework. The main idea about this framework is the application of multiagent architecture. As systems implementing Service Oriented Architecture are often geographically

and logically dispersed the appropriate tool for monitoring and controlling all components is necessary. The multiagent approach offer all relevant mechanisms and concepts so it seems to be the best solution in the described situation.

The main element of the multiagent system architecture for SOA security evaluation is definition of the agents classes. The following agents classes have been considered:

1. AMOL – SOA functional layer monitoring agents
2. ASL – SOA functional layer superior agents
3. AM – SOA managing agents
4. AC – the agents of consumers of SOA system services

Table 2. The characteristic of the agent classes

AMOL_1= $\{amol1_1, amol1_1, \dots, amol1_n\}$	<ul style="list-style-type: none"> - set of autonomous agents which perform the security level evaluation related tasks defined in Table 1. For the first functional layer of SOA – transport; - for example $amol1_1$ may be an agent that evaluates the confidentiality of the transport layer, $amol1_2$ may be an agent that evaluates data integrity at the transport layer level, etc.
AMOL_2= $\{amol2_1, \dots, amol2_m\}$ AMOL_3= $\{amol3_1, \dots, amol3_l\}$ AMOL_4= $\{amol4_1, \dots, amol4_o\}$ AMOL_5= $\{amol5_1, \dots, amol5_p\}$	<ul style="list-style-type: none"> - sets of autonomous agents for corresponding four SOA layers (communication protocols, ..., business processes) that perform specific security evaluation tasks related to the particular layer as described in Table 1.
ASL= $\{asl1, asl2, \dots, asl5\}$	<ul style="list-style-type: none"> - for each SOA functional layer there is defined one <i>superior agent</i>; - the <i>superior agents</i> range of responsibility is to coordinate all the tasks related to the security evaluation process for the particular SOA functional level, to collect the results provided by <i>amol</i> agents, to interpret the results provided by <i>amol</i> agents and finally to present the results of the security level to <i>managing agent</i> and to <i>client agents</i>
AM= $\{am\}$	<ul style="list-style-type: none"> - the <i>managing agent</i> is responsible for the most top-level security evaluation; it coordinate the activity of <i>asl</i> agents, collect the results of the SOA layer evaluation, combine all security level related information and produce the general SOA security level value, serves the <i>consumer agents</i> requests
AC= $\{ac_1, ac_2, \dots, ac_q\}$	<ul style="list-style-type: none"> - SOA <i>services consumers</i> agents collect the information about security level of provided by SOA systems services and evaluates the security level of composite services

The fundamental relations among agents, agent classes and SOA architecture are presented in fig. 3.

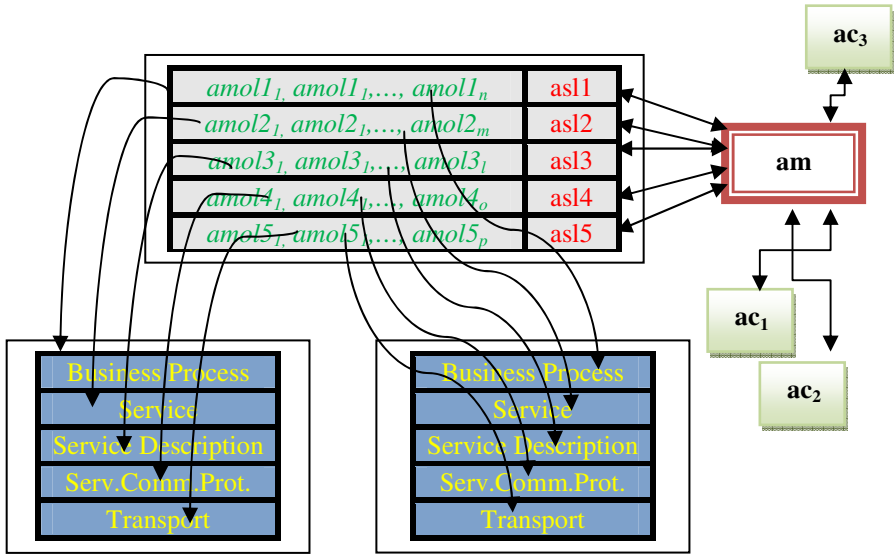


Fig. 3. The architecture of the multiagent system for SOA security level evaluation

4.2 A General Idea about SOA Security Level Evaluation

The most important functionality related to the SOA security level evaluation architecture is description of the all components, mechanisms and relations that are necessary to precisely evaluate the security level of the particular SOA system. As it was described in section 3 the problem of security evaluation is very complex and there exist more than one solution that could be acceptable within a context of a particular system and its environment. This part of the paper describes some general ideas about SOA security level evaluation in a relation to requirements listed in the table 1 and multiagent architecture presented in the fig. 3

Algorithm 1. Security level evaluation for a separate SOA functional layer.

Given:

- N – number of SOA layer
- ac_k – consumer of the service
- p_1, p_2, \dots, p_l – details or preferences related to ac_k request
- am – managing agent
- $amolN_1, \dots, amolN_m$ – set of specialized agents that perform security evaluation using appropriate tests, mechanisms, etc. related to the N -th SOA layer

Result:

- L_n – security level value for N -th layer

BEGIN

1. ac_k prepare and send to am a request concerning the security level of the N -th layer of the SOA system
2. am find all the monitoring agents related to N -th layer ($amolN_1, \dots, amolN_m$), prepare and send the appropriate requests to them
3. Monitoring agents ($amolN_1, \dots, amolN_m$) perform the security evaluation tasks using all tools, methods, algorithms, etc. available to them
4. am collects the results obtained by all monitoring agents and using the specific algorithm (data fusion, consensus operator, etc.) and taking into account the list p_1, p_2, \dots, p_l of ac_k preferences evaluates the final security level value of the N -th layer of this SOA system
5. am returns L_n to the ac_k

END

Algorithm 2. SOA security level evaluation.

- Given:**
- ac_k – consumer of the service
 - am – managing agent
 - p_1, p_2, \dots, p_l – details or preferences related to ac_k request
- Result:**
- L_{soa} – SOA system security level

BEGIN

1. ac_k prepare and send to am a request concerning the security level of the SOA system
2. Using *algorithm 1* managing agent am evaluate L_1, \dots, L_n – security levels for all SOA system's layers
3. Managing agent am evaluate L_{soa} the final security level value of the SOA system using selected data fusion methods and taking into account the list p_1, p_2, \dots, p_l of ac_k preferences
4. am returns L_{soa} to the ac_k

END

Discussion. In both algorithms there is no explicit definition of the method used for evaluation of the security level for a separate SOA layer and for the whole SOA system. The definition and the validation of the methods, algorithms used in these steps is one of the most challenging task of the security evaluation process. But as it was stated before, there is more than one acceptable approaches. The final decision concerning selection of the method used to combine the data provided by monitoring agents may depend on the context of the SOA system or/and the context of the consumer request.

5 Conclusion

The paper presents a novel framework of multiagent system for SOA security level evaluation. Also some general discussion about security level evaluation and Service Oriented Architecture have been presented. The multiagent architecture is composed of three types of agents: monitoring agents that tests the various security parameters

related to particular SOA layer, superior agents that manage the activity of monitoring agents, managing agents that are responsible for all superior agents and for communication with service consumer agents. Two algorithms used by monitoring agents and managing agents have been discussed.

The most important future work related to the problems described in the paper is proposition of the exact calculation method for assessment of the corresponding security level. After that, the selected method should be validated in the environment of the production SOA systems.

References

- [1] CERT (2009), <http://www.cert.org> (retrieved March 20, 2009)
- [2] Department of Homeland Security. National Vulnerability Database of the National Cybersecurity Division (2009), <http://nvd.nist.gov> (retrieved March 20, 2009)
- [3] Epstein, J., Matsumoto, S., McGraw, G.: Software security and SOA. *IEEE Security and Privacy* 4(1), 80–83 (2006)
- [4] Fernandez, E.B., Delessy, N.: Using patterns to understand and compare web services security products and standards (2006)
- [5] Kolaczek, G.: Opracowanie koncepcji specyfikacji metod i modeli szacowania poziomu bezpieczeństwa systemów SOA i SOKU, WUT (2009) (in polish)
- [6] Nakamura, Y., Tatsubori, M., Imamura, T., Ono, K.: Model-driven security based on web services security architecture. In: *IEEE International Conference on Services Computing*, vol. 1, pp. 7–15 (2005)
- [7] SANS Institute (2006), <http://www.sans.org> (retrieved March 20, 2009)
- [8] Skalka, C., Wang, X.: Trust by verify: Authorization for web services. Paper presented in *ACM Workshop on Secure Web Services*, pp. 47–55 (2004)
- [9] SOA Reference Model Technical Committee. *A Reference Model for Service Oriented Architecture*, OASIS (2006)
- [10] Steel, C., Nagappan, R., Lai, R.: *Core security patterns: Best practices and strategies for J2EE, web services, and identity management*. Pearson, Upper Saddle River (2006)
- [11] Tari, Z., Bertok, P., Simic, D.: A dynamic label checking approach for information flow control in web services. *International Journal of Web Services Research* 3(1), 1–28 (2006)
- [12] *WS-security policy 1.2*, OASIS (2009)
- [13] Yuan, E., Tong, J.: Attributed based access control (ABAC) for web services. In: *IEEE International Conference on Web Services*, pp. 561–569 (2005)

Describing Evolutions of Multi-Agent Systems

Sergey Babenyshev and Vladimir Rybakov

Department of Computing and Mathematics,
Manchester Metropolitan University,
John Dalton Building, Chester Street, Manchester M1 5GD, U.K.
and
Institute of Mathematics, Siberian Federal University,
79 Svobodny Prospect, Krasnoyarsk, 660041, Russia
Sergey.Babenyshev@gmail.com, V.Rybakov@mmu.ac.uk

Abstract. This paper¹ focuses on the issue of the formal logical description of evolutions of multi-agent systems (MAS). By evolution of a MAS we mean the change of inner states of the combined MAS caused by interaction of participating agents. We introduce a general scheme of combining propositional modal languages and respective logics into a single language suitable for such descriptions. The method is based on the representation of multi-agent systems by Kripke-Hintikka models. The obtained description allows to study the question of verifiable specifications.

Keywords: multi-agent systems, multi-modal logics, decision algorithms, satisfiability, Kripke semantics.

1 Introduction

Numerous attempts at combining logics — the hybrid, fusion, product logics providing by far not complete list — were motivated largely by applications, which more and more often require formalisms to describe complex systems with multiple ontologies. There are two main approaches toward combining propositional logic, that assume full combination of signatures. One is product and another is fusion [1]. From the point of view of applicability to modeling evolutions of MAS, there are certain constraints inherent to both these methods. Basic fusions do not allow for interaction of modalities. Products assume that structural configuration, representing a snapshot of a multi-agent system, should be fixed once and for all. Even more challenging restriction of products is that they are undecidable even for quite simple constituent logics, like *S5* [2].

For modeling behavior of multi-agent systems we propose a general scheme of *cluster-based fusions* that partially combines features of both fusion and product approaches, while avoiding their above-mentioned shortcomings. In particular, unlike products, cluster-based fusions lead to decidable logics, that allows, at

¹ This research is supported by Engineering and Physical Sciences Research Council, U.K. (EP/F014406/1).

least in principle, to use them for specification and verification purposes. Secondly, unlike the methods based on basic fusions, cluster-based fusions allow for far more expressive descriptive languages. The main difference with the previously used methods (cf. [3,4]), is that every state of a MAS is assumed to be contained inside a time-cluster.

The paper is structured as follows: in the next section we briefly review fusions and products of Kripke frames and discuss how our approach relates to them. In Section 3, we present the result about decidability of logics obtained by our method.

2 Cluster-Based Fusions of Kripke Frames

Let us recall the basics of multi-modal Kripke semantics.

A *Kripke multi-modal frame* F of the *Kripke signature* $\langle R_1, \dots, R_m \rangle$ is a tuple $\langle W, R_1, \dots, R_m \rangle$, where $W \neq \emptyset$ is the set of *worlds* or *states*, and every R_i is a binary relation on the elements of W (i.e., each $R_i \subseteq W \times W$). The set W is also called the *universe* of F , and the R_i s are called *accessibility relations*.

Later on, we will be dealing mainly with two types of Kripke frames and Kripke signatures. The first type, uni-modal, is used to imitate the flow of time. The second, multi-modal, is used to represent particular configurations of multi-agent system at given moments of time. Combined together, they represent an evolution of a MAS as a change of structural configurations in (possibly non-deterministic) time. Both type of Kripke frames will be assumed to satisfy certain structural conditions, that can be defined by modal formulas.

Let $F = \langle W, R, R_1, \dots, R_m \rangle$ be a multi-modal Kripke frame, where R is a reflexive, transitive relation. A *R-cluster* of F is a maximal under inclusion subset C of W , such that the restriction of R to C is an equivalence relation. We denote by $Cl_R(F)$ the set of all R -clusters of F . If the relation R is clear from context, we routinely drop the related subscripts or qualifications. Usually in this paper, the first relation in a combined Kripke signature will represent time, therefore we call the respective clusters — *time clusters*. The union of all clusters covers the frame and distinct clusters do not intersect. Therefore for every $w \in W$, we can define by $C(w)$ the unique cluster of F that contains w . In particular, any cluster of F is a cluster of the type $C(w)$ for some $w \in W$.

The main construction of this paper is given by

Definition 1. A multi-modal frame $\langle W, R, R_1, \dots, R_m \rangle$ is called a *cluster-based fusion (CB-fusion)* if

- R is a reflexive, transitive binary relation on W (R represents time);
- R_i are arbitrary relations on W (each R_i represents informational channels available to the i -th agent);
- each $R_i \subseteq R$ (i.e., all informational transactions are aligned with the time);
- each $R_i \subseteq R^{-1}$ (i.e., all informational transactions are determined by the current state of the MAS in question).

A *Kripke model* with the signature $\langle R_1, \dots, R_m \rangle$ is a tuple $\langle W, R_1, \dots, R_m, V \rangle$, where $\langle W, R_1, \dots, R_m \rangle$ is a Kripke frame with the signature $\langle R_1, \dots, R_m \rangle$, and $V : X \rightarrow \mathcal{P}(W)$ is a *valuation* of some subset $X \subseteq \text{Var}$ of variables. For a valuation V , $V(x_i)$ denotes the set of all worlds of W , where the basic fact, represented by the propositional variable x_i , is true.

For describing properties of multi-modal Kripke frames of the signature $\langle R_1, \dots, R_m \rangle$ the following modal propositional language is usually used:

$$\Lambda = \langle \wedge, \vee, \rightarrow, \neg, \diamond_1, \dots, \diamond_m \rangle.$$

We fix an enumerable set $\text{Var} := \{x_1, x_2, x_3, \dots\}$ of *propositional* variables. *Well-formed formulas of the language Λ (Λ -formulas)* are defined by the following grammar

$$\alpha ::= x_i \mid \alpha_1 \wedge \alpha_2 \mid \alpha_1 \vee \alpha_2 \mid \alpha_1 \rightarrow \alpha_2 \mid \neg \alpha \mid \diamond_i \alpha \quad (i \in I)$$

The set of all Λ -formulas is denoted by For_Λ . For a formula α , the *set of variables* $\text{Var}(\alpha)$ of α is defined inductively:

$$\text{Var}(x_i) := \{x_i\}, \quad \text{Var}(*\beta) := \text{Var}(\beta), \quad \text{Var}(\beta * \gamma) := \text{Var}(\beta) \cup \text{Var}(\gamma) \quad .$$

We will use the following shortcuts: $\square_i \phi := \neg \diamond_i \neg \phi$, $\top := p \vee \neg p$, $\perp := p \wedge \neg p$.

By Λ we denote the *standard* multi-modal language $\langle \wedge, \vee, \rightarrow, \neg, \diamond_1, \dots, \diamond_m \rangle$, where all \diamond_i are assumed to have the *standard* interpretation, i.e., given a Kripke model \mathcal{M} of the signature $\langle R_1, \dots, R_m \rangle$, for all $\phi \in \text{For}_\Lambda$, $w \in W$ and $i \in I$

$$(\mathcal{M}, w) \models \diamond_i \phi \Leftrightarrow \exists u : w R_i u \ \& \ (\mathcal{M}, u) \models \phi,$$

where we write $(\mathcal{M}, w) \models \alpha$ to say that the formula α is *true* or *holds* in the model \mathcal{M} at the world w . We will call the standard modalities (or their duals: $\square_i := \neg \diamond_i \neg$) with the presumed standard interpretation, *Kripke modalities*.

A formula α is *valid in a frame F* , if, for any valuation V of variables $\text{Var}(\alpha)$, $F \models_V \alpha$. If a formula α is not valid on F , then there is a valuation V such that $F \not\models_V \alpha$. In that case we say that α is *refuted* in F (by V).

Suppose \mathcal{K} is a class of Kripke frames of $\langle R_1, \dots, R_m \rangle$ -signature. Let Λ be the respective modal language. We denote by $\text{Log}(\mathcal{K})$ the logic

$$\{\alpha \in \text{For}_\Lambda \mid \mathcal{K} \models \alpha\}.$$

If $\mathcal{L} = \text{Log}(\mathcal{K})$ we say that \mathcal{L} is *generated* by \mathcal{K} .

If we have a multi-modal logic \mathcal{L} , let $\text{Fr}(\mathcal{L})$ be the class of all frames F of the respective Kripke signature, such that all theorems of \mathcal{L} are valid in F . A frame from $\text{Fr}(\mathcal{L})$ we will call *adequate for \mathcal{L}* or an \mathcal{L} -*frame*.

The simplest way of combining logics is the *fusion*. If \mathcal{L} is a Kripke complete unary modal logic, we denote \mathcal{L}_N the N -*fusion* of \mathcal{L} , i.e. the logic generated by the class of frames \mathcal{K} such that for every frame $F = \langle W, R_1, \dots, R_N \rangle \in \mathcal{K}$ and all $i \in \{1, \dots, N\}$, the frame $\langle W, R_i \rangle$ is an \mathcal{L} -frame.

The *product* $\mathcal{L}_1 \times \mathcal{L}_2$ of logics \mathcal{L}_1 and \mathcal{L}_2 is generated by *products of frames*. The product of $\langle W_1, R_1 \rangle$ and $\langle W_2, R_2 \rangle$ is the frame $\langle W_1 \times W_2, R_h, R_v \rangle$, where

(i) $\langle u_1, v_1 \rangle R_h \langle u_2, v_2 \rangle$ iff $(u_1 R_1 u_2$ and $v_1 = v_2)$; (ii) $\langle u_1, v_1 \rangle R_v \langle u_2, v_2 \rangle$ iff $(v_1 R_2 v_2$ and $u_1 = u_2)$. This construction can be iterated.

The combining scheme we propose is based on cluster-based fusions.

Definition 2. *Suppose we have two Kripke-complete logics \mathcal{B} and \mathcal{S} , of Kripke signatures $\langle R \rangle$ and $\langle R_1, \dots, R_m \rangle$, respectively, where \mathcal{B} is transitive. We call a $\mathcal{B}_\mathcal{S}$ -frame, every frame $F = \langle W, R, R_1, \dots, R_m \rangle$, such that (i) F is a CB-fusion, (ii) $\langle W, R \rangle$ is a \mathcal{B} -frame, (iii) for every $C \in Cl_R(F)$, $\langle C, R_1, \dots, R_m \rangle$ is an \mathcal{S} -frame.*

We define the cluster-based fusion of \mathcal{B} by \mathcal{S} , denoted by $\mathcal{B}_\mathcal{S}$, the logic

$$\{\alpha \in For_\Lambda \mid \text{for all } \mathcal{B}_\mathcal{S}\text{-frames } F : F \models \alpha\},$$

where $\Lambda = \langle \wedge, \vee, \rightarrow, \neg, \diamond, \diamond_1, \dots, \diamond_m \rangle$.

Let $\Lambda = \langle \neg, \wedge, \diamond_1, \dots, \diamond_m \rangle$ be a given modal language. A *clause* (or *type*) over variables x_1, \dots, x_n is a formula of the kind

$$\bigwedge_{k=1}^n x_k^{t(0,k)} \wedge \bigwedge_{i=1}^m \bigwedge_{k=1}^n (\diamond_i x_k)^{t(i,k)},$$

where x_i are variables, $t(i, k) \in \{0, 1\}$, and for any formula α , $\alpha^0 := \neg\alpha$, $\alpha^1 := \alpha$.

It is easy to see that there are only $2^{n(m+1)}$ distinct clauses over a set of n variables. We denote the set of all clauses over variables x_1, \dots, x_n by $\Theta(x_1, \dots, x_n)$.

For every $\theta \in \Theta(X)$ and $i \in \{1, \dots, m\}$, we denote

$$\mu^i(\theta) := \{x_k \in X \mid t(i, k) = 1\}.$$

We say that a clause θ is *realized* in a model \mathcal{M} , if there is a world $w \in \mathcal{M}$, such that $(\mathcal{M}, w) \models \theta$.

Given a Kripke model \mathcal{M} of the signature $\langle R_1, \dots, R_m \rangle$ with the finite *dom* V , every world $w \in W$ generates a unique clause $\theta_{\mathcal{M}}(w) \in \Theta(\text{dom } V)$, defined as

$$\theta_{\mathcal{M}}(w) := \bigwedge_{(\mathcal{M}, w) \models x_k^{t(0,k)}} x_k^{t(0,k)} \wedge \bigwedge_{(\mathcal{M}, w) \models (\diamond_i x_k)^{t(i,k)}} (\diamond_i x_k)^{t(i,k)},$$

where all $t(i, k) \in \{0, 1\}$. We will omit the subscript in $\theta_{\mathcal{M}}(w)$, whenever the model is clear from context. Also, we write for every $i \in \{0, 1, \dots, m\}$

$$\mu^i(\mathcal{M}) := \bigcup_{w \in W} \mu^i(\theta_{\mathcal{M}}(w)).$$

In particular, $\mu^0(\mathcal{M})$ is the set of the variables, that hold at least at one world in the model \mathcal{M} .

Definition 3. *Suppose \mathcal{M}_1 and \mathcal{M}_2 are two Kripke models of the same signature, such that \mathcal{M}_1 is finite and $\text{dom } V_1 = \text{dom } V_2$ is finite. Let $\Psi \subseteq \Phi \subseteq \Theta(\text{dom } V_1)$ be a pair of sets of clauses. We say that a mapping $f : W_1 \rightarrow W_2$ is a clause-preserving mapping of \mathcal{M}_1 into \mathcal{M}_2 modulo $\langle \Phi, \Psi \rangle$ if*

1. $\mathcal{M}_1 \models \bigvee \Phi$ and $\mathcal{M}_2 \models \bigvee \Phi$;
2. $\forall \theta \in \Psi \exists a \in \mathcal{M}_1 : (\mathcal{M}_1, a) \models \theta$;
3. $\forall \theta \in \Phi \forall a \in \mathcal{M}_1 : (\mathcal{M}_1, a) \models \theta \iff (\mathcal{M}_2, f(a)) \models \theta$;
4. $\mu^0(\mathcal{M}_1) = \mu^0(\mathcal{M}_2)$.

Thus the clause-preserving mapping $f : \mathcal{M}_1 \rightarrow \mathcal{M}_2$ modulo $\langle \Phi, \Psi \rangle$ guarantees that:

1. $\mathcal{M}_1, \mathcal{M}_2 \models \bigvee \Phi$,
2. all $\theta \in \Psi$ are realized in \mathcal{M}_1 ,
3. f preserves validity of all clauses from Φ .

Definition 4. Let \mathcal{L} be a Kripke complete logic and \mathcal{K} is a class of frames such that $\mathcal{L} = \text{Log}(\mathcal{K})$. We say that \mathcal{L} is complete under clause-preserving mapping w.r.t class \mathcal{K} if for every model \mathcal{M} over a frame $F \in \mathcal{K}$, and any sets of clauses $\Psi \subseteq \Phi \subseteq \Theta(\text{dom } V)$, there exists a finite model \mathcal{M}' such that

1. the frame of \mathcal{M}' is an \mathcal{L} -frame,
2. there exists a clause-preserving mapping of \mathcal{M}' into \mathcal{M} modulo $\langle \Phi, \Psi \rangle$.

If, in addition, each model \mathcal{M}' can be chosen so its size does not exceed $f(|\Phi|)$, for some computable function $f : \mathbb{N} \rightarrow \mathbb{N}$, then \mathcal{L} is said to be complete under f -bounded clause-preserving mappings w.r.t class \mathcal{K} .

3 Decidability Results

Following [5], we will be representing formulas by inference rules.

An (*inference*) rule is a pair $\langle \alpha, \beta \rangle$ of Λ -formulas. We will usually write the rule $\langle \alpha, \beta \rangle$ in the form α/β . For a rule $r = \alpha/\beta$: $\text{Var}(r) := \text{Var}(\alpha) \cup \text{Var}(\beta)$. A rule $r = \alpha/\beta$ is *valid in a model* \mathcal{M} (written $\mathcal{M} \models r$), if $\text{Var}(r) \subseteq \text{dom}(V)$ and

$$\mathcal{M} \models \alpha \implies \mathcal{M} \models \beta.$$

A rule r is *valid in a frame* F , if, for any valuation V of variables $\text{Var}(r)$, $F \models_V r$. If the rule r is not valid on F , then there is a valuation V such that $F \not\models_V r$. In that case we say that r is *refuted* on F (by V).

A rule r over the modal language $\Lambda = \langle \neg, \wedge, \vee, \rightarrow, \diamond_1, \dots, \diamond_m \rangle$ is said to be in the *reduced normal form* if

$$r = \bigvee_{1 \leq j \leq s} \theta_j / x_1, \quad (1)$$

and each disjunct θ_j has the form

$$\theta_j = \bigwedge_{k=1}^n x_k^{t(0,k)} \wedge \bigwedge_{i=1}^m \bigwedge_{k=1}^n (\diamond_i x_k)^{t(i,k)},$$

where x_i are variables, $t(i, k) \in \{0, 1\}$, and for any formula α , $\alpha^0 := \neg\alpha$, $\alpha^1 := \alpha$. Note that every disjunct of the reduced form is a clause.

Two rules r_1, r_2 are *equivalent* over a Kripke class \mathcal{K} , if for every $F \in \mathcal{K}$:

$$F \models r_1 \iff F \models r_2 .$$

For a formula α , the *set of subformulas* $Sub(\alpha)$ of α is defined as usually. For a rule $r = \alpha/\beta$: $Sub(r) := Sub(\alpha) \cup Sub(\beta)$.

It has been shown in Rybakov [6] that any modal inference rule may be transformed to an equivalent rule in the reduced normal form. Using essentially the same technique we can transform to normal reduced forms all rules of the considered modal language.

Lemma 1. *Let $\Lambda = \langle \neg, \wedge, \vee, \rightarrow, f_1, \dots, f_m \rangle$ be a language, where f_i are unary connectives (may be non-Kripke modalities). Suppose that for all f_i s holds*

$$f_i(p \leftrightarrow q) \leftrightarrow (f_i p \leftrightarrow f_i q).$$

Then any rule $r = \alpha/\beta$ can be transformed in exponential time to an equivalent rule r_{nf} in the reduced normal form.

From the definition of a normal reduced form, it is clear that under any given valuation of variables only one θ_j can hold true at a given state.

Thus, we have for every Λ -formula α and every frame F of the respective signature:

$$F \models \alpha \iff F \models x \rightarrow x/\alpha \iff F \models (x \rightarrow x/\alpha)_{\text{nf}} .$$

Therefore the following lemma holds:

Lemma 2. *A formula α is a theorem of a logic \mathcal{L} iff the rule $(x \rightarrow x/\alpha)_{\text{nf}}$ is valid in all \mathcal{L} -frames.*

Further on, rules will always be of the form (II).

Lemma 3. *Suppose \mathcal{B} and \mathcal{S} are two Kripke complete modal logics and \mathcal{S} is closed under f -bounded clause-preserving mappings. Then, if a rule $r = \bigvee_{1 \leq j \leq s} \theta_j/x_1$ is refuted on a $\mathcal{B}_{\mathcal{S}}$ -model, then r is refuted on a $\mathcal{B}_{\mathcal{S}}$ -model with the size of R -clusters at most $f(s)$.*

Definition 5. *Suppose \mathcal{M} is a transitive Kripke model of the signature $\langle R \rangle$, such that $\text{dom } V$ is finite. A model \mathcal{N} is a clause-filtration of the model \mathcal{M} if*

$$W = W/\sim, \text{ where } u \sim v \iff \theta_{\mathcal{M}}(u) = \theta_{\mathcal{M}}(v).$$

$$[u]_{\sim} R_i [v]_{\sim} \iff \mu^i(\theta_{\mathcal{N}}(v)) \subseteq \mu^i(\theta_{\mathcal{N}}(u)),$$

$$V(x_k) = \{[w]_{\sim} \mid x_k \in \mu^0(\theta_{\mathcal{N}}(w))\}.$$

We say that a Kripke logic \mathcal{L} *admits strong clause-filtration*, whenever for every \mathcal{L} -model \mathcal{M} over an \mathcal{L} -frame, there is a clausal filtration \mathcal{N} of \mathcal{M} , based on an \mathcal{L} -frame.

The strong clause-filtration property is a variant of the usual filtration property [7,8,9], modified in two respects: firstly, it adjusted for the use with the reduced normal forms, secondly, it requires the existence of a filtration model with underlying frame adequate for \mathcal{L} .

Lemma 4. *Let $\mathcal{M} = \langle W, R, R_1, \dots, R_m, V \rangle$ be the \mathcal{B}_S -model obtained in Lemma 3, in particular $\mathcal{M} \not\models r$. Suppose also that the logic \mathcal{B} is closed under clause-filtrations. Then r is refuted on a finite \mathcal{B}_S -model of the size less or equal than $f(s) \cdot 2^s$, where $s = |\Theta(r)|$.*

A Kripke logic \mathcal{L} has the *finite model property*, whenever for every formula $\alpha \notin \mathcal{L}$, there is a finite model \mathcal{M} such that $\mathcal{M} \models \mathcal{L}$ and $\mathcal{M} \not\models \alpha$. If, in addition, the model \mathcal{M} can be chosen to be of the size not more than $f(|\alpha|)$, for some computable function $f : \mathbb{N} \rightarrow \mathbb{N}$, then \mathcal{L} has the *f-effective finite model property*.

We say that a Kripke-complete logic \mathcal{L} *admits strong filtration*, whenever for every model $\mathcal{M} \not\models \alpha$ over an \mathcal{L} -frame, there is a finite set of formulas Σ , such that there is a model $\langle F, V \rangle \not\models \alpha$, that is a filtration of \mathcal{M} modulo Σ and $F \in Fr(\mathcal{L})$. The logics that admit strong filtration form the majority of standard logics to which the *filtration method* (see [10]) can be applied. They include $K4$, $S4$, $S5$ and so on. A logic \mathcal{L} that admits strong filtration modulo sets of the kind $Sub(\alpha)$, also admits strong clausal filtration.

Theorem 1. *Let*

1. \mathcal{B} be a transitive logic that admits strong filtration,
2. \mathcal{S} be a multi-modal logic, closed under f -bounded clause preserving mappings,

Then \mathcal{B}_S has the g -effective finite model property, where $g(x) = (f(x) + 1) \cdot 2^x$.

Proof. By Lemmas 3 and 4 □

Corollary 1. *If under conditions of Theorem 1 the class of finite models of \mathcal{B}_S is decidable (i.e., the set of isomorphic classes of finite \mathcal{B}_S -models is decidable), then \mathcal{B}_S is decidable.*

As an easy corollary of Theorem 1 we obtain:

Corollary 2. *Suppose \mathcal{B} is a mono-modal transitive logic that admits strong filtration. If \mathcal{S} also admits strong filtration, then \mathcal{B}_S has the effective finite model property. In particular, if*

$$\mathcal{B} \in \{K4, S4, S5\}, \quad \mathcal{S} \in \{K4_N, S4_N, S5_N\},$$

then the logic \mathcal{B}_S has the effective finite model property.

Proof. Since the filtration property implies the clause-filtration property, therefore, by Theorem 1, the logic \mathcal{B}_S has the effective finite model property □

Corollary 3. *If, in addition to conditions of Corollary 2, logics \mathcal{B} and \mathcal{S} have effectively recognizable classes of finite models, then \mathcal{B}_S is decidable.*

Proof. Since, by Corollary 2, logic \mathcal{B}_S has the finite model property, it suffices to show that the class of finite \mathcal{B}_S -frames is effectively recognizable. To check that a given frame $F = \langle W, R, R_1, \dots, R_m \rangle$ is a \mathcal{B}_S -frame we only need to check that

1. the frame $\langle W, R \rangle$ is a \mathcal{B} -frame,
2. the frame $\langle C, R_1, \dots, R_m \rangle$ is \mathcal{S} -frame, for every $C \in Cl_R(F)$.

Both conditions can be checked effectively, and also the procedure of recognizing clusters in a finite frame is effective. Thus logic \mathcal{B}_S is decidable. □

4 Conclusion and Future Work

The paper presents a method for describing evolutions of MAS, which differs from the established approaches in several important respects. First of all, unlike the methods based on basic fusions, it allows for more expressive descriptive language. On the other hand, we prove that, unlike products of logics used as a base for describing evolutions of MAS, the proposed cluster-based approach leads to decidable logics. We demonstrate this by presenting a generic decision algorithm. This algorithm has 2EXPTIME-complexity (relative) bound, which makes it practically unfeasible. Nevertheless, the decidability of corresponding logics, opens a possibility for obtaining more practical variants of the decision algorithms, possibly using tableaux-based techniques.

References

1. Caleiro, C., Carnielli, W., Rasga, J., Sernadas, C.: Fibring of logics as a universal construction. In: Gabbay, D., Guenther, F. (eds.) *Combination of Logics. Handbook of Philosophical Logic*, vol. 13. Kluwer, Dordrecht (2005)
2. Gabbay, D.M., Kurucz, A., Wolter, F., Zakharyashev, M.: *Many-dimensional modal logics: theory and applications. Studies in Logic*, vol. 148. Elsevier Science, Amsterdam (2003)
3. Fagin, R., Halpern, J., Moses, Y., Vardi, M.: *Reasoning About Knowledge*. MIT Press, Cambridge (1995)
4. Halpern, J.Y., van der Meyden, R., Vardi, M.Y.: Complete axiomatizations for reasoning about knowledge and time. *SIAM Journal on Computing* 33(3), 674–703 (2004)
5. Rybakov, V.: Logical consecutions in intransitive temporal linear logic of finite intervals. *Journal of Logic Computation* 15(5), 633–657 (2005)
6. Rybakov, V.: A criterion for admissibility of rules in the modal system S4 and the intuitionistic logic. *Algebra and Logica* 23(5), 369–384 (1984)
7. Segerberg, K.: Decidability of S4.1. *Theoria* 34, 7–20 (1968)
8. Gabbay, D.M.: Selective filtration in modal logic. *Theoria* 36, 323–330 (1970)
9. Lemmon, E., Scott, D.: *An Introduction to Modal Logic*. Blackwell, Oxford (1977)
10. Chagrov, A., Zakharyashev, M.: *Modal Logic. Oxford Logic Guides*, vol. 35. Clarendon Press, Oxford (1997)

Functionality and Performance Issues in an Agent-Based Software Deployment Framework

Mario Kusek, Kresimir Jurasovic, and Ignac Lovrek

University of Zagreb
Faculty of Electrical Engineering and Computing
Unska 3, HR-10000 Zagreb
{mario.kusek,kresimir.jurasovic,ignac.lovrek}@fer.hr

Abstract. Deploying and maintaining software in a distributed system includes software delivery, remote installation, starting, stopping, and modifying in order to configure or re-configure a system according to user needs. This paper deals with an agent-based framework where intelligent and mobile agents provide the means to implement a distributed system and enable its evolution by taking partial or full responsibility for software deployment tasks. Agents are organised into agent teams, where one agent is the team leader responsible for planning, while the others are operational agents capable of executing a defined plan. The formal model, as well as functionality and performance issues, are elaborated. Special attention is paid to deployment strategies and their optimization, while taking into account characteristics of distributed system nodes and the network connecting them. Simulation-based evaluation of agent serialization, migration and deserialization parameters, and their influence on overall performance, is included.

1 Introduction

Establishment and maintenance of distributed systems includes operations that provide software delivery to system nodes, remote installation, starting and stopping, and version handling. Furthermore, modification of software after delivery must be supported in order to correct faults, improve performance characteristics, adapt to a changed environment or improve maintainability. Distributed system software should be configured initially, and re-configured if and when corrective, perfective, adaptive or preventive actions are required. Software deployment and maintenance strategies are important in order to configure or re-configure distributed systems according to their requirements (i.e., where and what), and to achieve minimum configuration and re-configuration setup times (i.e., when and how), taking into account the network and node characteristics, as well as operational conditions.

Intelligent and mobile agents provide the means to implement a distributed system and enable its evolution, by taking partial or full responsibility for these resource intensive and costly tasks. Regarding system performance, configuration and re-configuration should have minimum influence on system operation.

Consequently, software deployment and maintenance should be optimised in order to achieve acceptable total execution time. In this paper, functionality and performance issues of an agent-based framework are elaborated. Furthermore, system parameters related to agent migration and agent activation/deactivation, and their influence on overall performance, are discussed.

The paper is organized as follows: Section 2 describes formal model of an agent-based software deployment framework and introduces a prototype system MA-RMS. Case study dealing with system parameters describing agent serialization/de-serialization time and communication link capacity, including results of simulations is presented in Section 3. Section 4 concludes the paper.

2 The Formal Model of an Agent-Based Software Deployment Framework

Software deployment has to take into account the functionality that a distributed system provides (i.e. services), as well as characteristics of nodes (i.e. capacity, operating system, agent platform, installed software) and the network connecting them (i.e. topology, link bandwidth). Furthermore, procedures specific to the system must be considered. Examples of distributed systems faced with such problems are network-centric applications and networks themselves, with hundreds or thousands of nodes, e.g. access points in local networks or base stations in mobile networks [1,2].

From the formal standpoint, a distributed system is defined by the tuple (S, N) where S denotes system nodes, $S = \{S_1, S_2, \dots, S_i, \dots, S_{ns}\}$, and N denotes a network connecting them. System nodes should be configured in order to provide support for a set of required elementary services, $ES = \{es_1, es_2, \dots, es_j, \dots, es_{nes}\}$ provided by the system, i.e., a defined subset of elementary services $s_i = \{es_{i1}, es_{i2}, \dots, es_{ij}, \dots, es_{in}\}$ should be supported by each node S_i . Distributed system configuration is defined with an initial set-up where software components corresponding to elementary services from ES are delivered and activated at each node S_i , according to its predefined functionality, s_i . Re-configuration is required when the ES changes (i.e. new elementary service is introduced or an existing one updated), when node S changes (i.e. new node is connected or an existing one functionally upgraded) or when network N changes (i.e. network topology or link capacity changes).

An agent-based software deployment system is organised as a multi-agent system, A_{SD} , which shares the set of nodes, S and the network N with the system under consideration, (A_{SD}, S, N) . A_{SD} includes a planning agent a_P and a team of operational agents, $\{a_1, a_2, \dots, a_i, \dots, a_{na}\}$. The planning agent a_P is responsible for planning software deployment, allocating deployment tasks to operational agents in the team and co-ordinating them. Operational agents are multi-operational, capable of executing one or more deployment tasks to one or more nodes. Each individual task corresponds to an operation, such as software delivery to a system node, remote installation, starting or stopping, replacement/modification of the existing software, version handling and others.

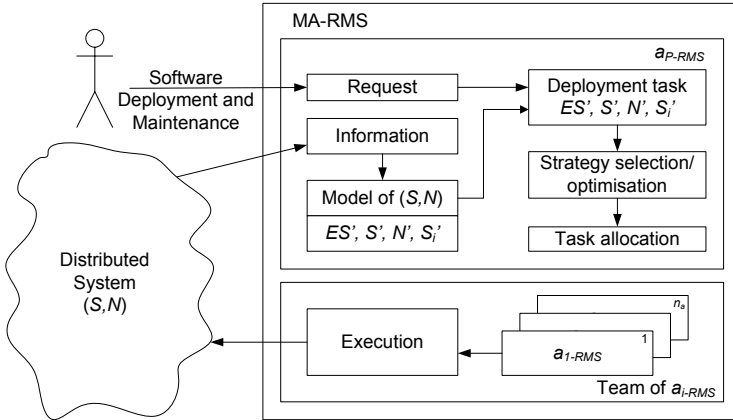


Fig. 1. Service deployment framework MA-RMS

When deploying software components on a large number of nodes at remote locations using software agents, one of the major challenges faced is determining the number of operational agents to use and distributing the required tasks among them. Software deployment efficiency depends on agent team organisation and the strategy used. The simplest strategy is when a single agent executes all tasks on all nodes, while the most promising one is when team size and task assignment are optimized according to the system and network conditions. As an example, the role of planning agent a_P , when ES should be re-configured to ES' ($ES \rightarrow ES'$), is defined as follows: (1) Identify a subset of nodes $S' \subset S$ that should be re-configured, i.e. for which $s'_i \neq s_i$; (2) Define deployment tasks required for re-configuration of each node $S_i \in S'$, for each elementary service $es_{ij} \in s'_i$; (3) Define a sub-network $N' \subset N$ which connects nodes from S' ; (4) Collect node and network parameters for (S', N') ; (5) Decide which deployment strategy to use, organize the agent team and let them work; (6) If deployment results with ES' , all tasks were completed successfully. Otherwise the resulting configuration is $ES^x \neq ES'$, and the procedure for re-configuring $ES^x \rightarrow ES'$ should be repeated until completion.

Related work includes different aspects of the agent-based approach for software (re)configuration, updating and maintenance [3,4,5]. Using agents in service oriented architectures has been studied mostly for service composition and orchestration [6]. The practical implementation can be difficult for complex distributed systems, from functionality and performance point of view. Some preliminary results show that rational agents based on BDI paradigm can cope with a complexity issues related to frequent changes, deployment errors or inconsistency and rollback ability. A service deployment framework, called the Multi-Agent Remote Maintenance Shell (MA-RMS) [4,7] shown in Fig. 1, is based on a formal model of a mobile agent network [8].

The main goal of the planning agent, a_{p-RMS} , is to optimize the deployment strategy. This goal is divided into two sub goals. The first sub-goal is to gather

and analyse information regarding the nodes and the network. The second sub-goal is to select a previously defined fixed strategy or an optimized one, taking into account the node and network characteristics and conditions. This goal consists of a set of plans, defined by triggering conditions and plan implementation. Triggering conditions define when the plan is applicable, while plan implementation activates operational agents according to the selected strategy. Planning agent a_{p-RMS} selects the predefined strategy which gives minimum completion time. The strategies are the following: (R1) a single agent executes all services on all nodes; (R2) an agent executes a single service on one node only; (R3) an agent executes all services on one node only; (R4) an agent executes a specific service on all nodes; (R5) services are assigned to the agents in order to exploit maximal parallelism in service execution (mutually independent services are assigned to different agents, in order to execute them simultaneously); (R6) a hybrid solution combining R4 and R3. An agent is responsible for a specific service on all nodes while other agents execute the remaining services each on a different node. Strategies R1–R6 are static, i.e., they always generate the same distribution of tasks among agents for a specific network topology, regardless of available link bandwidth or other conditions. An additional strategy, R7, based on a genetic algorithm for agent team optimization uses the network topology and link bandwidth as input parameters [9].

Operational agents, a_{i-RMS} , support full functionality required for software deployment and maintenance on nodes in an IP network. These agents know how to migrate, install, configure, start, stop and uninstall software. A mobile agent network simulator is applied to access performance issues [10]. It can simulate agent execution in different networks (i.e. nodes, switches and links), different operation execution times and different strategies, including the genetic algorithm.

3 Evaluating Performance of the Agent-Based Software Deployment Solution

The basic performance measure evaluated in the paper is the total execution time of the software deployment and maintenance process. The first group of parameters influencing this process describe the software under consideration (ES) and its distribution to system nodes (S). This distribution is defined by software configuration requirements for each node (s_i). The second group of parameters is related to agents, their complexity (i.e. the operations assigned to them, their size) and their life-cycle, as shown in Fig. 2. Agent migration between nodes S_i and S_j when performing some deployment task is defined by an agent transfer time, as follows: $T_{ij} = t_{pi} + t_{ij} + t_{aj}$ where t_{pi} is the agent preparation time needed for agent serialization at the originating node S_i ; t_{ij} is communication time needed for agent transfer from S_i to S_j ; and t_{aj} is the agent activation time which includes agent reception and de-serialisation at the destination node.

Handling of some deployment task at node S_j is described by an agent holding time: $t_{qj} = t_{cj} + t_{wj} + t_{sj}$, where t_{cj} is the inter-agent communication time (i.e.

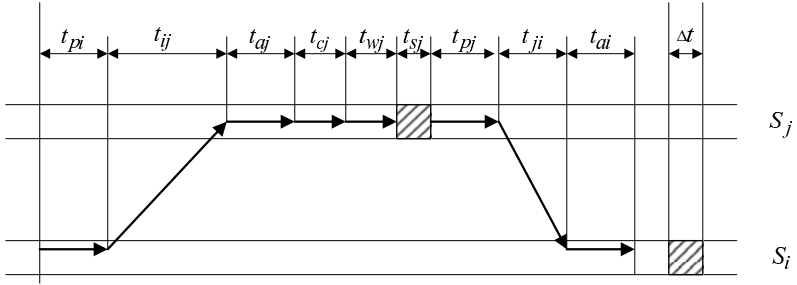


Fig. 2. Agent transfer and holding time

the time an agent spends at node S_j searching for the result of a deployment task performed by another agent); t_{wj} is waiting time (i.e. the time an agent spends in a queue at S_j waiting execution); and t_{sj} is the serving time (i.e. the time needed for execution at S_j). The third group of parameters describes the characteristics and conditions of a distributed system that affect software deployment and maintenance. The basic node characteristics taken into consideration are node processing power influencing serving time, and agent serialization and de-serialization (t_{sj} , t_{pi} , t_{aj}). All deployment tasks are treated the same with respect to t_{sj} , which equals Δt . The basic characteristics of the network include its topology and available link bandwidth. For analysis purposes, all parameters related to timing are expressed in discrete time units, Δt .

This approach provides a tool for analyzing the performance of a single agent. When extended to a multi-agent system and a mobile agent network [11], it allows for analysis of the total execution time required to fulfill a software deployment and maintenance request. The total execution time is the time required by an agent team to execute all deployment tasks on all nodes. The case study that follows deals with the impact of agent serialization/de-serialization (t_{pi} and t_{aj} on Fig. 2) and available link bandwidth on the total execution time. The sub-network under consideration consists of 12 nodes and 6 switch components connected with a link of certain bandwidth (Fig. 3). Node S0 is used as a starting node hosting the planning agents and from which all agents start their execution. The set of elementary services that should be deployed to all nodes except S0 consists of 4 services. Three network scenarios are simulated by varying two parameters: a) the available link bandwidth, and b) the ratio of agent execution time to agent serialization/de-serialization time. The aim of the simulation is to explore the influence of bandwidth variation and agent complexity on strategy selection. In the first scenario the bandwidth of the links in the network was set to 5 Mbit/s, in the second to 10 Mbit/s and in the third to 100 Mbit/s.

In all scenarios, the service agent execution time to agent serialization/de-serialization time ratio varies from 1 to the 128. The results of the simulations are shown in Fig. 4 (Sc. 1), Fig. 5 (Sc. 2) and Fig. 6 (Sc. 3). The graphs for all scenarios and all strategies show the same basic characteristic: growth of the total execution time as the ratio of agent execution time to agent serialization/

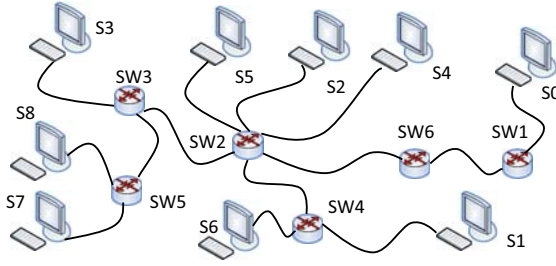


Fig. 3. Network topology

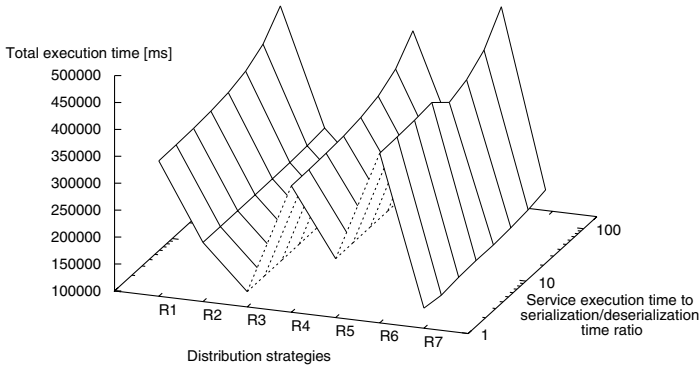


Fig. 4. Scenario 1: Simulation result for link bandwidth of 5 Mbit/s

de-serialization time increases. The reason for this is evident: holding time increases with agent complexity. The difference between results obtained from different strategies is the speed of its growth.

Each agent should be activated before execution and the strategy R3 with a single agent responsible for one node only is the best one in this respect. Such a conclusion will lead to the planning agent to define the triggering condition for R3 as “low bandwidth” & “wide execution time span” and make it applicable after detecting such situations (Sc. 1 and 2). Strategies R1, R4, and R6 have poor performance in slow networks because of intensive agent migration, while strategies R3 and R7 give acceptable total execution times (Sc. 1 and 2). With higher bandwidth available, strategies R2 and R6 become comparable to R3, because migration time has far less impact due to fast links (Sc. 3). Furthermore, faster networks compensate with higher agent ratio, making complex agents more attractive. Strategy R5 requires a triggering condition “parallel capability”. Optimized strategy R7 shows good performance in the worst conditions, i.e. in the case when the available bandwidth is low (Sc. 1 and 2). This discussion covers performance analysis assuming an ideal environment where software deployment and maintenance is completed as required. In reality, some agents might not

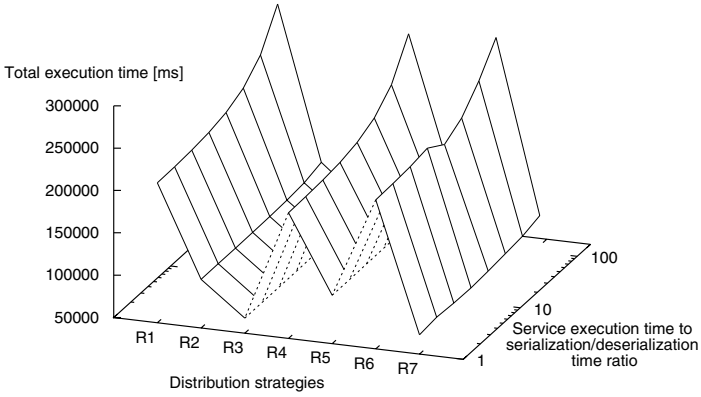


Fig. 5. Scenario 2: Simulation result for link bandwidth of 10 Mbit/s

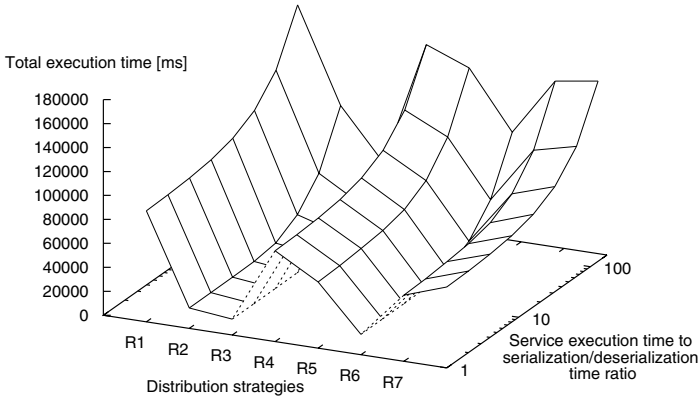


Fig. 6. Scenario 3: Simulation result for link bandwidth of 100 Mbit/s

perform as required due to network congestion, unavailable nodes, and errors or faults, leading to partial fulfilment of software deployment tasks. When detected, such events can be used as additional triggers for optimized strategies.

4 Conclusion

An agent-based framework for software deployment includes a multi-agent system with a planning agent and a team of operational agents. The intelligent planning agent needs to know the current state of system nodes and the network. Using this information, it can decide which strategy to choose in order to obtain better performance characteristics, i.e. acceptable total execution time. Simulations were performed to evaluate deployment strategies from the performance perspective. Three network scenarios were simulated with variations in the link bandwidth and

the ratio of agent execution time to agent serialization/de-serialization time. The obtained results demonstrate how bandwidth variation and agent complexity influence the total execution time.

Future work will include further definition and implementation of the planning agent as a rational agent based on the BDI paradigm. Further on, robustness issues will be studied in order to exploit the full potential of an agent-based software deployment framework in a large-scale environment.

Acknowledgments

This work is part of research project 036-0362027-1639 "Content Delivery and Mobility of Users and Services in New Generation Networks", supported by the Ministry of Science, Education and Sports of the Republic of Croatia.

References

1. Sherif, M.H., Ho, S.: Evolution of operation support systems in public data networks. In: IEEE Symposium on Computers and Communications, p. 72 (2000)
2. Houssos, N., Alonistioti, A., Merakos, L., Mohyeldin, E., Dillinger, M., Fahrmaier, M., Schoenmakers, M.: Advanced adaptability and profile management framework for the support of flexible mobile service provision. Special Issue on (R)Evolution towards 4G Mobile Communication Systems 10 (2003)
3. Bettini, L., De Nicola, R., Loreti, M.: Software Update via Mobile Agent Based Programming. In: Proceedings of SAC, Special Track on Agents, Interactions, Mobility, and Systems, pp. 32–36. ACM Press, New York (2002)
4. Jezic, G., Kusek, M., Desic, S., Caric, A., Huljenic, D.: Multi-agent remote maintenance shell for remote software operations. In: Grid Services Engineering and Management. LNCS (LNAI), vol. 2774, pp. 675–682. Springer, Heidelberg (2003)
5. Dalpiaz, F., Giorgini, P., Mylopoulos, J.: Software self-reconfiguration: a bdi-based approach. In: Proceedings of the 12th The Eight International Conference on Autonomous Agents and Multiagent Systems (2009)
6. Ventakesan, V., Portchlevic, V.: Architecture for services orchestration using bdi agent, <http://msdn.microsoft.com/en-us/library/bb898865.aspx> (10.03.2009)
7. Lovrek, I., Caric, A., Huljenic, D.: Remote maintenance shell: Software operations using mobile agents. In: Proceedings of the International Conference on Telecommunications, pp. 175–179 (2002)
8. Sinkovic, V., Lovrek, I.: Generic model of a mobile agent network suitable for performance evaluation. In: KES, pp. 675–678 (2000)
9. Jurasovic, K., Kusek, M.: Optimizing service distributions using a genetic algorithm. In: Lovrek, I., Howlett, R.J., Jain, L.C. (eds.) KES 2008, Part I. LNCS (LNAI), vol. 5177, pp. 158–165. Springer, Heidelberg (2008)
10. Kusek, M., Jurasovic, K., Jezic, G.: Verification of the mobile agent network simulator - a tool for simulating multi-agent systems. International Journal of Software Engineering and Knowledge Engineering 18, 651–682 (2008)
11. Sinkovic, V., Kusek, M., Jezic, G., Lovrek, I.: Performance evaluation of a mobile agent network using network calculus. In: Lovrek, I., Howlett, R.J., Jain, L.C. (eds.) KES 2008, Part I. LNCS (LNAI), vol. 5177, pp. 174–181. Springer, Heidelberg (2008)

A Consensus-Based Integration Method for Security Rules

Trong Hieu Tran and Ngoc Thanh Nguyen

Institute of Informatics, Wroclaw University of Technology, Poland
Trong.Hieu.Tran@pwr.wroc.pl, Ngoc-Thanh.Nguyen@pwr.wroc.pl

Abstract. Policy-based security is an effective approach to manage knowledge systems by handling all behaviors of a system thought a set of rules. This approach has such advantages as capacity to define general high-level targets, ease for configuration, and flexibility in development and maintenance. However, the resolution of conflicts is unavoidable requirement because of many elements of subjectivity as well as objectivity in administrative processes. To this end, several works have been done, and they gave concrete results. In this paper, we will propose a new approach to solve conflicts and to integrate rules in a policy. A new representation of rules is given, the distances between rules are defined as well as postulates are presented and analyzed. Algorithms for integrating policy also have been proposed and examined.

Keywords: Policy-based security, knowledge integration.

1 Introduction

Security is one of the important problems in multi-agent systems as well as database systems. It has become increasingly important for computer and information systems with explosive growth of the Internet and the widespread use of wireless networks. There are some methods to ensure the security for a system, and policy-based management is one of the most common and effective approaches. By this approach, we can set the configuration easily, have the flexibility in development and maintenance processes as well as define general tasks in a high level without knowing about the detailed specification of system in which policy is applied. This approach is applied in sub-fields of AI and database systems such as security and access management [9], network management and monitoring [6, 12], and electronic commerce [2, 7].

In policy-based management approach, all behaviors of a system are handled by a sequence of rules called policy. As we mentioned above, the approach has some advantages. However, if an inconsistent situation arises, it may lead the system to an unknown state or an error. Unfortunately, this situation is difficult to avoid because the rules of a policy may be given by many administrators, in different periods of time, and without the clear idea of their purposes [1]. Therefore, the problem to integrate security policy can be formulated as follows:

Given policy may contain some conflicts, which requires resolution for the policy to be a robust one.

Working out a solution of this problem is one of the basic requirements of the system administrating.

In order to integrate policies, the common task that we have to perform is solving conflicts. There are several methods for conflict resolution proposed. In [13], authors propose the method base on Consensus Theory to resolve conflicts and integrate security policy. The methods, based on the order of rules in the policy, the priority of the restriction of rules, and the most/least specific condition, are introduced in [11]. Some structures defined for representing policies have been also examined. In [1] the authors defined an algebra of security policy as well as its semantics to combine authorization specifications. The hierarchy structures are used in [3, 4] and graph is recommended in [5, 10]. The representation of security rules on the syntactic level has been surveyed and analyzed in works published recently. For instance, set-based approach and semi-lattices are used to solve conflicts in policy rules [1] or relational structures have been proposed in [13]. In this paper, we propose the approach to represent this kind of knowledge and to solve conflicts on a logical semantic level.

The rest of this paper is structured as follows. In Section 2, we present some related concepts such as some definitions of rule, policy, and conflict. The distance functions between rules are introduced in Section 3. The postulates are proposed and some algorithms are examined in Section 4. At last, some conclusions are included in Section 5.

2 Basic notions

Definition 1. A *rule* is the binding of a condition with an action to handle the behavior of the security system at a concrete situation. The condition will be evaluated to determine whether the action is performed.

Formally, the model of policy rule is presented as follows:

$$r: C \rightarrow a$$

in which C is a family of conditions, a is an action, and symbol “ \rightarrow ” is only the way to represent the concept of the binding of a condition with an action instead of usual logical meaning.

The condition usually is understood as specific fields of values, the number of these fields in each system is usually constant, and they are strictly ordered. The condition is satisfied if and only if all the fields are satisfied. In this work, we also assume that the conditions of rules include finite fields, which are ordered in a definite order and the actions of rules belong to a definite set of actions defined by administrators.

Definition 2. A *policy* is a sequence of rules, which is used to administer, manage, and control access to a security system [1].

A policy determines the appropriate action, which will be performed for each particular situation of system. So, a policy is also understood as the information that can be

used to handle the behavior of a system. Obviously, the system may have changed behavior if the order of rules in its policy is changed. Formally, we denote a policy p including rules r_1, r_2, \dots , and r_n by a sequence as follows:

$$p = \langle r_1, r_2, \dots, r_n \rangle$$

By symbol “ \oplus ” we denote the concatenation between two policies, so a policy can be built as follows:

1. If r is a rule, $\langle r \rangle$ is a policy.
2. If p_1 and p_2 are policies, $p_1 \oplus p_2$ is also a policy.

It is easy to notice that with p_1, p_2 and p_3 are policies; the concatenating operator \oplus has following characteristics:

- a) $p_1 \oplus p_2 \neq p_2 \oplus p_1$
- b) $(p_1 \oplus p_2) \oplus p_3 = p_1 \oplus (p_2 \oplus p_3)$

According to semantic view, a policy p can be defined as a pair $\langle C, A \rangle$ where C is a set of fields of conditions and A is a set of actions. Policy p includes a set of rules, in which each of them has a conjunction of condition fields in C called a condition and an action in set A . We also accept the assumption that each rule has exactly one action. Therefore, we concentrate about the representation of conditions of rules as follows:

The real world of conditions includes a set $C = \{c_1, c_2, \dots, c_n\}$ of fields of conditions and a set $V = \{V_{c_1}, V_{c_2}, \dots, V_{c_n}\}$ of the elementary values of condition fields respectively, (each V_{c_i} is the set of values of condition field c_i , or V_{c_i} is super domain of c_i). Shortly, pair (C, V) is called a real world of conditions. Let $\prod(V_c)$ denote the set of all subsets of set V_c . We also assume that for each condition field c , its value is always a set of elementary values from V_c , and obviously, it is an element of set $\prod(V_c)$. An *elementary value* means a value, which is indivisible in the system.

An expression $(c = v)$ or $(c \neq v)$ where $c \in C, v \in \prod(V_c)$ and v is a finite set, is called a *literal* from real world (C, V) . If a literal has form $(c = v)$ we call it a *positive literal* if it has form $(c \neq v)$ then we call a *negative literal*. A negative literal $(c \neq v)$ can be considered to be equivalent to $\neg(c = v)$. A negative literal may be transformed into a positive literal by using the attribute super domains, that is literal $(c \neq v)$ is equivalent to literal $(c = v')$ where $v' = V_c \setminus v$. By C_{CV} we denote the set of all conditions of (C, V) -based literals.

Definition 3. By the semantics of conditions, we define the following function:

$$S_{Co}: C_{CV} \rightarrow \prod_{c \in C} V_c$$

such that

$$S_{Co}(x) = \{(a_1, a_2, \dots, a_n): a_i \in v_i, i=1..n\}$$

where $x = (c_1, v_1) \wedge (c_2, v_2) \wedge \dots \wedge (c_n, v_n), v_i \in \prod(V_{c_i})$

Thus the semantics of condition x is the set of all tuples built by Cartesian product of all super domains of the condition fields occurred in x . The intuition of this definition is based on the aspect that if condition x represents the condition of a rule, set $S_{Co}(x)$ will consist of all possible scenarios which are included in x .

Definition 4. *The semantics of rules is the semantics of conditions binding with the corresponding actions.*

Because of the assumption that each rule has only one action, it is intuitive to consider that the semantics of a rule includes all possible scenarios of the condition binding with the action. The following example illustrates the intuition:

Example 1. Considering to a rule in access filter policy of a system as follows: $r = (protocol, \{TCP\}) \wedge (IP_address, \{192.168.0.2-4\}) \wedge (port, \{100-102\}) \rightarrow (action = permitted)$.

The semantics of rule r has the following tuples:

Table 1. The semantics of rule r

Protocol	IP_address	Port	Action
TCP	192.168.0.2	100	Permitted
TCP	192.168.0.2	101	Permitted
TCP	192.168.0.2	102	Permitted
TCP	192.168.0.3	100	Permitted
TCP	192.168.0.3	101	Permitted
TCP	192.168.0.3	102	Permitted
TCP	192.168.0.4	100	Permitted
TCP	192.168.0.4	101	Permitted
TCP	192.168.0.4	102	Permitted

We have the following properties of the semantics of conditions:

Proposition 1. *Conditions*

$$x = (c_1, v_1) \wedge (c_2, v_2) \wedge \dots \wedge (c_k, v_k) \text{ and} \\ x' = (c_1, v_1) \wedge (c_2, v_2) \wedge \dots \wedge (c_k, v_k) \wedge (c, V_c)$$

where attribute c does not occur in x , should have the same semantics, that is

$$S_{Co}(x) = S_{Co}(x').$$

Conditions x and x' having the same semantics are called *equivalent* to each other.

Definition 5. *Rules $c_1 \rightarrow a_1$ and $c_2 \rightarrow a_2$ are conflict if $S_{Co}(c_1) \cap S_{Co}(c_2) \neq \emptyset$.*

The conflict between two rules occurs in the case if it there exists scenarios in which the rules have the same condition. In work [13], authors classified and analyzed types of policy conflicts based on the relations between rules such as shadowing conflict, redundancy conflict, correlation conflict, and exception conflict.

3 Distances between Security Rules

Generally, the distance between two rules may be understood as the sum of the distance between the conditions and the distance between the actions of these rules. It is intuitive that the distance between two conditions should be equal the minimal cost of translating the semantics of the first condition into the semantics of the second one. Thus we have:

Definition 6. For conditions $b = c_1 \wedge c_2 \wedge \dots \wedge c_n$ and $b' = c'_1 \wedge c'_2 \wedge \dots \wedge c'_m$, their distance $d_c(b, b')$ is equal the minimal cost for transforming set $Sc(b)$ into set $Sc(b')$.

By the operation transforming set $Sc(x)$ into set $Sc(x')$ we mean performing such operations as *adding*, *removing* and *transformation* to the elements of set $Sc(x)$, which in the result give set $Sc(x')$. For the need of the definition of these operations, we define the following cost functions:

- Function $d_1: V \rightarrow (0, +\infty)$: specifies the cost for adding (or removing) of an elementary value to (or from) a set.
- Function $d_2: V \times V \rightarrow [0, +\infty)$: specifies the cost for transformation of one elementary value into another.

Similarly, like in work [8] for functions d_1 and d_2 we also accept the following assumptions:

a) Function d_2 is a metric, i.e. for any $x, y, z \in V$ the following conditions are held:

- $d_2(x, y) \geq 0$, $d_2(x, y) = 0$ if and only if $x=y$,
- $d_2(x, y) = d_2(y, x)$,
- $d_2(x, y) + d_2(y, z) \geq d_2(x, z)$;

b) For any $x, y \in V$ $|d_1(x) - d_1(y)| \leq d_2(x, y) \leq d_1(x) + d_1(y)$.

Condition a) ensure that function d_2 may be treated as a distance function between elements from set V . Condition b) is an intuitive condition, that may be stated that from set $\{y\}$, the cost to add element x into $\{y\}$ to become $\{x, y\}$ should be smaller than the cost to transform y into x and add y to become $\{x, y\}$; the cost to transform x into y is smaller than the total cost to remove x and to add y .

For convenience in calculating, in this work we assume that $d(x) = d(y) = 1$ and $d(x, y) = d(x) + d(y)$.

Definition 7. For rules $r_1 = c_1 \rightarrow a_1$ and $r_2 = c_2 \rightarrow a_2$, the distances between r_1 and r_2 is calculated as

$$d(r_1, r_2) = d_C(c_1, c_2) + d_A(a_1, a_2)$$

$$\text{where } d_A(x, y) = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{if } x \neq y \end{cases} \quad (x, y \in A)$$

We have the algorithm to calculate the distance between two rules as follows:

Algorithm 1. Computing distance value of two rules.

Given: Rules $r_1 = c_1 \rightarrow a_1$ and $r_2 = c_2 \rightarrow a_2$

Result: Distance value between r_1 and r_2 .

BEGIN

1. For each $c \in C$, and V_c is super domain of c .

If c occurs in c_1 but not in c_2 then $c_2 = c_2 \wedge (c, V_c)$

If c occurs in c_2 but not in c_1 then $c_1 = c_1 \wedge (c, V_c)$

2. For each condition field c in c_1 , we calculate the cost to transform (c, v_1) in c_1 into (c, v_2) in c_2 .

3. Calculate the cost to transform action a_1 into action a_2 .

4. Return total value of *Step 2* and *Step 3*

END

4 Postulates and Algorithms

Let U be a finite universe consisting of rules may occur in a policy system. By $\prod(U)$ we denote the set of all finite and nonempty sequences with repetitions of set U . Each element of $\prod(U)$ is called a *conflict profile* in policy system. Therefore, a conflict profile is a set with repetition of rules with a determined order, in other words, it is called a policy profile in some system. By integration function, we mean the following function:

$$\mathcal{C} : \prod(U) \rightarrow 2^U$$

In this function, we assume that the result will be a sequence without repetition. For a profile X sequence $\mathcal{C}(X)$ is called the integration of X . By $\mathcal{C}(U)$ we denote the set of all integration functions for universe U .

Definition 8. By an integration function $\mathcal{C} \in \mathcal{C}(U)$ for profiles of rules we understand a function:

$$\mathcal{C} : \prod(U) \rightarrow 2^U$$

which satisfies one or more of the following postulates:

P1. For $X = \langle x_1, x_2, \dots, x_n \rangle$, $x_i \in U$, there should be:

$$\mathcal{C}(X \oplus X \oplus \dots \oplus X) = \mathcal{C}(X)$$

P2. For $X = X_1 \oplus X_2$ and Y is a subsequence of $\mathcal{C}(X)$, there should be

$$\mathcal{C}(X) = \mathcal{C}(X_1 \oplus Y \oplus X_2)$$

P3. $\mathcal{C}(X) \neq \emptyset$ for any profiles X

P4. For $X = X_1 \oplus X_2$, there should be

$$\mathcal{C}(X) = \mathcal{C}(\mathcal{C}(X_1) \oplus \mathcal{C}(X_2))$$

P5. A consensus $x^* \in \mathcal{C}(X)$ should minimize the sum of distances:

$$\sum_{x \in X} d(x^*, x) = \min_{x' \in X} \sum_{x \in X} d(x', x);$$

P6. an O_2 -consensus of set P if it satisfies the following condition:

$$\sum_{x \in X} d^2(x^*, x) = \min_{x' \in X} \sum_{x \in X} d^2(x', x)$$

Some commentary of these postulates is given as follows:

- Postulate $P1$ corresponds with situation in which policy includes some identical subsequences of rules. The result of integrating process is the policy will be changed by a subsequence.
- Postulate $P2$ states that if a subsequence of rules is in the integration sequence of the profile X , it is also in the integration sequence of profile created by insert this subsequence into X .
- Postulate $P3$ implies that there is always solution for any integration process.
- The main idea of postulate $P4$ is that we can first partition the policy into some subsequences of rules and perform integration processes for each subsequence, and then we concatenate these intermediate results and perform the last integration process to achieve result. The idea is based on divide-and-conquer strategy, a very common one in Artificial Intelligent.
- Postulates $P5$ and $P6$ refer to the popular criteria of consensus theory. We can use these criteria to determine the integration result quantitatively.

With the assumption that all condition fields are independent and all actions are independent, we present algorithm to integrate policy rules based on $O1$ criterion as follows:

Algorithm 2. Computing $O1$ -consensus X for set P of rules.

Given: Finite set P (with repetitions) of (C, A) -based clauses.

Result: Consensus X for P satisfies $P1$, $P2$, and $P5$.

BEGIN

1. Create the set:

$$profile(P) = \bigcup_{r \in P} S(r)$$

where $S(x)$ is the set of semantics of rule x ;

2. Let $X := \emptyset$; $S_b := +\infty$; $Z = profile(P)$;

3. While $Z \neq \emptyset$ do

- 3.1. Select from Z an element z such that the sum

$$\min := \sum_{y \in profile(b)} \min_{x \in X \cup \{z\}} d(x, y) \text{ is minimal;}$$

$$Z := Z \setminus \{z\};$$

- 3.2. If $S_b \geq \min$ then

Begin

$$S_b := \min$$

$$X := X \cup \{z\};$$

End;

End while

4. Return X ;

END.

The idea of this algorithm is stated as follows: we firstly collect all the semantics of all rules in set P of rules, and then we step by step choose the best semantics, of which the distance among it and chosen ones is minimum and the reached value is smaller than current total distance value, and add it in chosen set of semantics. We perform these steps until all semantics in initial set is examined.

5 Conclusions

In this paper, we have proposed a new approach to solve conflicts and to integrate rules in a policy, in which a new representation of policy rules on the semantic level has been presented and the distances between rules are defined. Several postulates for policy integration are proposed and analyzed. An algorithm for policy integration has been proposed and examined. In future works, we will continue to have more deeply analysis about this approach and work out some algorithms satisfying other groups of chosen postulates.

References

- [1] Cataldo, B., Antonio, L.: Algebraic Models to Detect and Solve Policy Conflicts. In: MMM - ACNS 2007, pp. 242–247 (2007)
- [2] Grosf, B.N., Labrou, Y., Chan, H.Y.: A Declarative Approach to Business Rules in Contracts: Courteous Logic Programs in XML. In: Wellman, M.P. (ed.) Proc. First ACM Conf. Electronic Commerce, pp. 68–77 (1999)
- [3] Guoli, D., Jianhua, C., Robert, F.L., Peter, P.C.: Graph-theoretic method for merging security system specifications. *Inf. Sci.* 177(10), 2152–2166 (2007)
- [4] Kagal, L., Finin, T., Joshi, A.: A policy based approach to security for the semantic web. In: Fensel, D., Sycara, K., Mylopoulos, J. (eds.) ISWC 2003. LNCS, vol. 2870, pp. 402–418. Springer, Heidelberg (2003)
- [5] Koch, M., Mancini, L.V., Parisi-Presicce, F.: Administrative scope in the graph-based framework. In: Proceedings of SACMAT 2004, Yorktown Heights, NY, pp. 97–104 (2004)
- [6] Hasan, M.Z.: An Active Temporal Model for Network Management Databases. In: Proc. IFIP/IEEE Fourth Int'l Symp. Integrated Network Management, pp. 524–535 (1995)
- [7] Minsky, N.H., Ungureanu, V.: A Mechanism for Establishing Policies for Electronic Commerce. In: Proc. 18th Int'l Conf. Distributed Computing Systems, pp. 322–331 (1998)
- [8] Nguyen, N.T.: Consensus System for Solving Conflicts in Distributed Systems. *Information Sciences – An International Journal* 147, 91–122 (2002)
- [9] Kiyohiko, O., Nariyoshi, Y., Hayato, I., Kota, A., Toshio, M.: An Efficient Management Method of Access Policies for Hierarchical Virtual Private Networks. In: Proceeding of COMSWARE, Bangalore, India, pp. 1–7 (2007)
- [10] Sandhu, R.S.: Role-based access control. *Advances in Computers*, vol. 48, pp. 237–286. Academic Press, London (1998)
- [11] Castano, S., Fugini, M., Martella, G., Samarati, P.: *Database Security*. Addison Wesley, Reading (1994)
- [12] Davy, S., Jennings, B., Strassner, J.: Efficient Policy Conflict Analysis for Autonomic Network Management. In: Fifth IEEE Workshop on Engineering of Autonomic and Autonomous Systems, EASE 2008, pp. 16–24 (2008)
- [13] Tran, T.H., Nguyen, N.T.: Security Policy Integration Method for Information Systems. In: Proceedings of ACIISD 2009, pp. 223–225. IEEE CS Press, Los Alamitos (2009)

Emotion Judgment Based on Relationship between Speaker and Sentential Actor

Seiji Tsuchiya¹, Eriko Yoshimura¹, Fuji Ren², and Hirokazu Watabe¹

¹ Dept. of Intelligent Information Engineering and Sciences, Doshisha University
Kyo-Tanabe, Kyoto, 610-0394, Japan

² Institute of Technology and Science, The University of Tokushima
Minami Josanjima, Tokushima, 770-8506, Japan
{stsuchiy,eyoshimu,hwatabe}@mail.doshisha.ac.jp,
ren@is.tokushima-u.ac.jp

Abstract. Authors are conducting research aiming to develop new interfaces that follow the mechanism of human communication, particularly focusing on human common sense. In this paper, a method is proposed which processes any "subject" using knowledge base and an Association Mechanism. In proposed method, 27 attributes of "subject" were judged by knowledge base. Moreover, an unknown word processing is proposed which deals with actor words which were not registered in the knowledge base. The result of the proposed method gave the correct answer in 75% of cases. If the "not out-of-common-sense" answers were counted as part of the "correct answers", the correct-answer ratio rose to 96%. Therefore, if the proposed method and the existing method were combined, the correct-answer ratio was approximately 85%.

Keywords: Emotion, Common Sense, Concept Base, Degree of Association.

1 Introduction

Authors are conducting research aiming to develop new interfaces that follow the mechanism of human communication, focusing on human common sense. Humans, in such communication, are able to appropriately interpret ambiguous information that they receive and carry on a smooth conversation. Common sense is knowledge (ability) that only man has. The person can express, and act feeling neither sense of incompatibility nor unnatural by using common sense. Moreover, when the sense of incompatibility and unnatural are felt, the person can appropriately interpret them.

Especially, authors focus on the emotion of such common sense and attempt to establish a method to judge the user's feelings based on what the user says. It is expected that use of this system can, for instance, select an appropriate expression if the content that the system tries to provide the user contains expressions that are unpleasant or remind the user of unhappy events.

Such systems and methods have already been developed. The developed method [1] judges a user's emotion, categorized into 10 types, from a sentence the

user utters, based on the four components of the sentence: "subject", "modifier", "object word", and "action word". However, "subject" used in the method has been limited to "I".

However, for example, people judge that speaker is joyful from utterance "My father obtains a lot of money". On the other hand, people judge that speaker is angry from utterance "Thief obtains a lot of money". Thus, proper processing of sentential actor is absolutely imperative for a smooth conversation. Therefore, a method is proposed which processes any "subject" using knowledge base and an Association Mechanism in this paper.

2 The Existing Emotion Judgment System

The components of uttered sentences to be used to judge speaker emotion were limited to four ("subjects", "modifiers", "objects" and "action words") [1]. Figure 1 shows outline of the existing Emotion Judgment method.

A "subject" was a noun that refers to the agent of the uttered sentence. This was limited to "I" which denotes the speaker him/herself.

A "modifier" was an adjective or "adjectival verb" that modifier the "object" which follows the modifier. "Modifiers" may be omitted, as they were not always necessary in textual expression. The direct modification and dependent modification types were further divided into different groups having similar meaning according to the adjectives describing the modifiers, and they were registered in the knowledge base for emotion judgment.

An "object" was a noun that denotes the object of the subject's action, behavior, or state. Objects were also classified according to their meanings using the 203 sense words that the Sense Judgment System [2, 3] can judge. These 203 sense words share the common meaning categories with the modifiers discussed earlier. In addition, "modifiers" and "objects" collectively were referred to as "object words". In short, the 203 sense words were used to categorize the meanings of the object words.

An "action word" was a verb, adjective, or "adjectival verb" that describes the subject's action, behavior, or state. An action word converted the feature

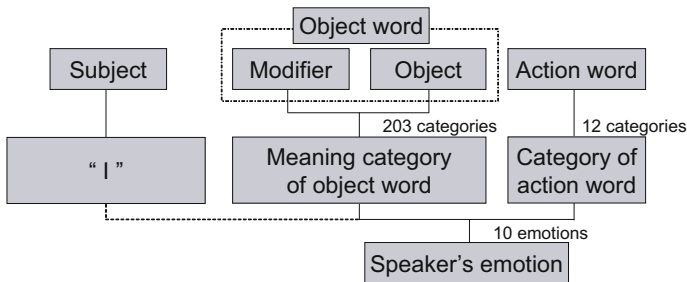


Fig. 1. Outline of the existing Emotion Judgment System

related to the sense and perception that associated with an object word. Features expressed in terms of senses and perceptions could be roughly divided into positive and negative expressions. Likewise, emotions could also be categorized into two groups, positive and negative. Therefore, four types of effect could be found in the action words.

A speaker's emotion was judged based on the "subject", "object words", and "action words". With respect to the emotions that were generated, those associated with a total of 406 pairs of the meaning categories of object words (203 categories) and action words (2 categories of "succession" and "opposite") were manually defined and registered in the system's Emotion Judgment Knowledge Base.

Many psychologists, philosophers, etc., have studied human emotions [4, 5, 6]. However, these researchers have had different interpretations of emotions and devised different models for emotions, as emotions have no substance and are quite ambiguous. Therefore, emotions have been defined as "something one feels instantaneously when an action takes place" and has defined the following ten emotions to judge: "joy", "sadness", "anger", "ease", "fear", "disappointment", "shame", "regret", "sense of guilt", and "no emotion".

Some knowledge related to the "generation of emotions", the "action words", and the "modifiers" of the "object words" were registered in the Emotion Judgment Knowledge Base. Based on this, the system associated words and expanded its knowledge within the range of common sense, making it possible to handle many expressions. The word association was realized by using the huge Concept Base [7, 8] that was automatically built from multiple digital dictionaries, and a method to calculate the Degree of Association [9] that evaluates the relationship between words. Hereafter, this Concept Base and the calculation method are called the "Association Mechanism".

3 Elemental Technique

3.1 Concept Base

The Concept Base is a large-scale database that is constructed both manually and automatically using words from multiple electronic dictionaries as concepts and independent words in the explanations under the entry words as concept attributes. In the present research, a Concept Base containing approximately 90,000 concepts was used, in which auto-refining processing was carried out after the base had been manually constructed. In this processing, attributes considered inappropriate from the standpoint of human sensibility were deleted and necessary attributes were added.

In the Concept Base, Concept A is expressed by Attributes a_i indicating the features and meaning of the concept in relation to a Weight w_i denoting how important an Attribute a_i is in expressing the meaning of Concept A . Assuming that the number of attributes of Concept A is N , Concept A is expressed as indicated below. Here, the Attributes a_i are called Primary Attributes.

$$A = \{(a_1, w_1), (a_2, w_2), \dots, (a_N, w_N)\}$$

↑ Concept	train, 0.36	locomotive, 0.21	railroad, 0.10	...	a_1, w_1	} Primary Attributes
	train, 0.36	locomotive, 0.21	railroad, 0.10	...	a_{i1}, w_{i1}	
	locomotive, 0.21	streetcar, 0.23	subway, 0.25	...	a_{i2}, w_{i2}	} Secondary Attributes
	⋮	⋮	⋮	⋮	⋮	
	a_{1j}, w_{1j}	a_{2j}, w_{2j}	a_{3j}, w_{3j}	...	a_{ij}, w_{ij}	

Fig. 2. Example of the Concept “train” expanded as far as Secondary Attributes

Because Primary Attributes a_i of Concept A are taken as the concepts defined in the Concept Base, attributes can be similarly elucidated from a_i . The Attributes a_{ij} of a_i are called Secondary Attributes of Concept A . Figure 1 shows the elements of the Concept ”train” expanded as far as Secondary Attributes.

3.2 Degree of Association Algorithm

For Concepts A and B with Primary Attributes a_i and b_i and Weights u_i and v_j , if the numbers of attributes are L and M , respectively ($L \leq M$), the concepts can be expressed as follows:

$$A = ((a_1, u_1), (a_2, u_2), \dots, (a_L, u_L))$$

$$B = ((b_1, v_1), (b_2, v_2), \dots, (b_M, v_M))$$

The Degree of Identity $I(A, B)$ between Concepts A and B is defined as follows (the sum of the weights of the various concepts is normalized to 1):

$$I(A, B) = \sum_{a_i=b_j} \min(u_i, v_j)$$

The Degree of Association is calculated by calculating the Degree of Identity for all of the targeted Primary Attribute combinations and then determining the correspondence between Primary Attributes. Specifically, priority is given to determining the correspondence between matching Primary Attributes. For Primary Attributes that do not match, the correspondence between Primary Attributes is determined so as to maximize the total degree of matching. Using the degree of matching, it is possible to give consideration to the Degree of Association even for Primary Attributes that do not match perfectly.

When the correspondences are thus determined, the Degree of Association $R(A, B)$ between Concepts A and B is as follows:

$$R(A, B) = \sum_{i=1}^L I(a_i, b_{xi})(u_i + v_{xi}) \times \{\min(u_i, v_{xi}) / \max(u_i, v_{xi})\} / 2$$

In other words, the Degree of Association is proportional to the Degree of Identity of the corresponding Primary Attributes, and the average of the weights of those attributes and the weight ratios.

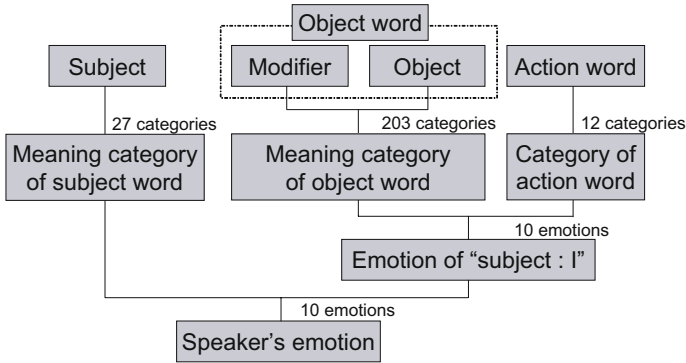


Fig. 3. Outline of proposed Emotion Judgment System

3.3 Sense Judgment System [2, 3]

The knowledge base for the sense and perception judgments has a structure like a thesaurus, and it contains sense and perception words that are associated with typical nouns, which have been entered manually. In cases when an unknown word not registered in the Sense Knowledge Base needs to be processed, the system calculates the Degree of Association with those known words registered in the knowledge base for the sense and perception judgments and chooses the one with the highest Degree of Association for processing. This lets the system obtain the rough corresponding sense and perception. In addition, the system refers to the attributes registered in the Concept Base to find the sense and perception particular to that word. Due to its structure, these attributes in the Concept Base contain some inappropriate words as senses and perceptions to be associated, and thus the system is carefully designed so that the correct sense and perception is selected using the Degree of Association.

4 Processing of Emotion Judgment Based on the Relationship between Speaker and Sentential Actor

The existing Emotion Judgment System has been limited to "I" which denotes the speaker him/herself for "subject". However, for example, people judge that speaker is joyful from utterance "My father obtains a lot of money". On the other hand, people judge that speaker is angry from utterance "Thief obtains a lot of money". Thus, proper processing of sentential actor is absolutely imperative for a smooth conversation. Therefore, an emotion judgment method is proposed which any "subject" is processable in this paper.

Figure 3 shows outline of the proposed Emotion Judgment System.

"Object word" and "action word" are categorized into 203 sense words and 12 categories, respectively. Moreover, emotion of "subject(I)" are judged by these combination. The method is same in section 2.

In the proposed method, to correspond besides "I" which is "subject" in the existing method, the processing of the subject is enhanced. Concretely, "subject" is categorized into 3 attributes: liking (likes and dislikes), familiarity (closeness), sociality (good and evil). These 3 attributes have 3 values. In short, "subject" is categorized into 27 categories. Speaker's emotion is judged by combination of these 27 categories and judged emotion of "I". In addition, 3 attributes of "subject" and speaker's emotion are judged by knowledge base.

4.1 Sentential Actor's Attribute Knowledge

A knowledge base for judgment of sentential actor's emotion was manually defined. The knowledge base was created based on an existing thesaurus [10] using a tree structure to represent its knowledge efficiently. All nouns registered in the thesaurus were related to the above-mentioned 3 attributes values. In addition, the nouns related to the 3 attributes values were 9068 words.

4.2 Unknown Word Processing for Sentential Actor Judgment

Even if the attributes values of the sentential actor are related to 9068 words by using the thesaurus, all sentential actors cannot be covered. Thus, an unknown word processing is proposed which deals with words which were not registered in the knowledge base.

As mentioned, words which are lower than some node have similar attributes values, because the knowledge base was created based on the thesaurus. Therefore, when input word are not registered in the knowledge base, the Degree of Association are calculated between the word and 437 nodes (words) in the thesaurus. Moreover, Words which have the Degree of Association among the top ten are selected. Attributes values of unknown word are defined by a majority vote of the attributes values of the selected words. As a result, the attributes values can be related to the unknown word.

4.3 Knowledge for Emotion Judgment

As mentioned, speaker's emotion is judged by combination of attributes values of sentential actor and judged emotion of "I". Therefore, an emotion generation knowledge base was newly made besides the existing emotion generation knowledge base. 270 rules which are combination of 27 categories (attributes values) of sentential actor and 10 emotions are registered in the knowledge base.

5 Performance Evaluation of the Proposed Emotion Judgment Method

In order to evaluate how valid the emotions generated by proposed emotion judgment method were, 530 sentences were collected from 5 test subjects to be used as data for evaluation. In addition, in this evaluation, "object word" ("modifier"

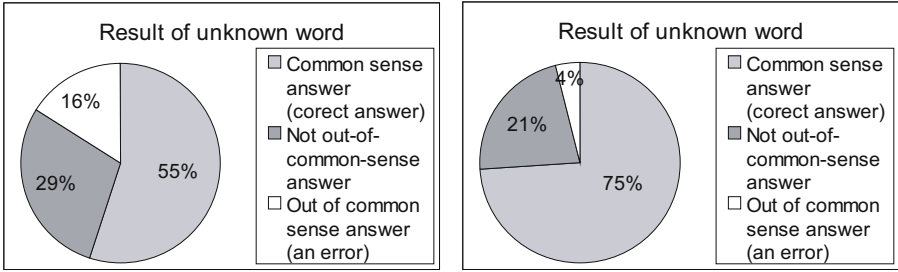


Fig. 4. Result of proposed Emotion Judgment method

and "object") and "action word" which were perfectly processed were selected for evaluation of sentential actor's processing. Five test subjects were then asked to judge whether the emotions generated by the proposed method were common sense. If four or more of judges said emotion generated was a common sense emotion, the emotion was considered as "common sense" (correct answer). In cases when two or three subjects said the emotion was "common sense", the generated emotion was considered as being "not out-of-common-sense". If only one subject or no subjects considered the generated emotion as "common sense", the emotion was thought of as "out of common sense" (an error). For cases when multiple emotions were generated, they were considered as being "common sense" (correct answer) if all of the generated emotions were common sense ones. If any one of the emotions was considered to be "out of common sense", then the particular emotion generated was considered as being "out of common sense" (an error). All others were regarded as being "not out-of-common-sense".

Figure 4 shows the result of the proposed emotion judgment method. Sentences which had unknown word of sentential actor were 80 sentences. The result gave the correct answer in 55% of cases. If the "not out-of-common-sense" answers were counted as part of the "correct answers", the correct-answer ratio rose to 84%. In addition, all result gave the correct answer in 75% of cases. If the "not out-of-common-sense" answers were counted as part of the "correct answers", the correct-answer ratio rose to 96%. Furthermore, the existing Emotion Judgment method had 88% as the correct-answer ratio. If the proposed method and the existing method were combined, the correct-answer ratio was approximately 85%. Because these ratios were high, authors believe that proposed method is effective.

6 Conclusions

In this paper, authors focused on emotions, which are part of the common sense judgments humans make in everyday communication. Such existing method judges a user's emotion, categorized into 10 types, from a sentence the user utters, based on the four components of the sentence: "subject", "modifier",

”object word”, and ”action word”. However, ”subject” used in the method has been limited to ”I”. Therefore, a method was proposed which processed any ”subject” using knowledge base and an Association Mechanism. In proposed method, 27 attributes of ”subject” were judged by knowledge base. Moreover, an unknown word processing is proposed which deals with actor words which were not registered in the knowledge base.

The result of the proposed method gave the correct answer in 75% of cases. If the ”not out-of-common-sense” answers were counted as part of the ”correct answers”, the correct-answer ratio rose to 96%. Therefore, if the proposed method and the existing method were combined, the correct-answer ratio was approximately 85%. Because these ratios were high, authors believe that proposed method is effective.

Acknowledgment

This research has been partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (Young Scientists (B), 21700241).

References

- [1] Tsuchiya, S., Yoshimura, E., Watabe, H., Kawaoka, T.: The Method of the Emotion Judgment Based on an Association Mechanism. *Journal of Natural Language Processing* 14(3), 119–238 (2007)
- [2] Horiguchi, A., Tsuchiya, S., Kojima, K., Watabe, H., Kawaoka, T.: Constructing a Sensuous Judgment System Based on Conceptual Processing. In: Gelbukh, A. (ed.) *CICLing 2002. LNCS*, vol. 2276, pp. 86–95. Springer, Heidelberg (2002)
- [3] Watabe, H., Horiguchi, A., Kawaoka, T.: A Sense Retrieving Method from a Noun for the Commonsense Feeling Judgement System. *Journal of Artificial Intelligence* 19(2), 73–82 (2004)
- [4] Hukui, Y.: *Kanjo-No-Shinrigaku* (Japanese). Kawashima Shoten (1990)
- [5] Rita, C.: *No-To-Kokoro-No-Chizu* (Japanese). Hara Shobo (1999)
- [6] Susan, G.: *No-No-Tankyu* (Japanese). Mumeisha (2001)
- [7] Hirose, T., Watabe, H., Kawaoka, T.: Automatic Refinement Method of Concept-base Considering the Rule between Concepts and Frequency of Appearance as an Attribute. Technical Report of the Institute of Electronics, Information and Communication Engineers. NLC2001-93, pp. 109–116 (2002)
- [8] Kojima, K., Watabe, H., Kawaoka, T.: A Method of a Concept-base Construction for an Association System: Deciding Attribute Weights Based on the Degree of Attribute Reliability. *Journal of Natural Language Processing* 9(5), 93–110 (2002)
- [9] Watabe, H., Kawaoka, T.: Measuring Degree of Association between Concepts for Commonsense Judgements. *Journal of Natural Language Processing* 8(2), 39–54 (2001)
- [10] NTT Communication Science Laboratory: *Nihongoitaikei*. Iwanami Shoten (1997)

A Knowledge Based Formal Language for Securing Information Systems

Yun Bai

Intelligent Systems Laboratory
School of Computing and Mathematics
University of Western Sydney
Locked Bag 1797, Penrith South DC
NSW 1797, Australia
ybai@scm.uws.edu.au

Abstract. In this paper, we propose a formal logic approach to specify the system security policies and rules and their reasoning in response to queries of accessing the system resource. Especially we investigate and handle the situation where the security agent's knowledge based on which the access decision is made is not complete. We introduce modal logic to specify and reason about a security domain, then translate the domain into an epistemic logic program [10]. We show that our approach has an expressive power to describe a variety of complex security scenarios.

Keywords: Authorization Policy, Modal Logic, Formal Language, System Security, Logic Program.

1 Introduction

Authorization or access control protects the information system by only allowing authorized entry and operation on the system resources. This topic has been extensively studied in [3,11] etc. and a variety of authorization specification approaches such as access matrix [5], role-based access control [4], access control in database systems [8], authorization delegation [9], procedural and logical specifications [12] have been investigated. Since logic based approaches provide a powerful expressiveness [6] as well as flexibility for capturing a variety of system security requirements, increasingly, a lot of work has been focusing on this aspect. Nevertheless, there were some limitations so far in these approaches. For instance, when the security agent does not have complete, specific information on the security domain, how to reason and answer access queries under such a scenario?

In this paper, we propose a knowledge based formal languages \mathcal{L}^k to specify authorization domain with incomplete information in secure computer systems. We introduce modal logic to specify and reason about a security domain then translate the domain into an epistemic logic program. We show that our approach has an expressive power to describe a variety of complex security scenarios.

To simplify our presentation, we assume the existence of a single, local system security officer or security agent administering the authorizations. This assumption

enables us to concentrate on a single administering agent system and hence avoids the problem of coordination among multi agents.

The rest of the paper is organized as follows. Section 2 describes language \mathcal{L}^k by outlining its syntax and gives some authorization policy examples specified by the language. Section 3 explains the semantics of language \mathcal{L}^k . We start by introducing a general overview of epistemic logic program, then map the domain description specified by \mathcal{L}^k into the logic program. In section 4, we show a case study of a domain description and its reasoning. Finally, section 5 concludes the paper with some remarks.

2 A Knowledge Based High Level Language \mathcal{L}^k

In this section we define the basic syntax of a high level language \mathcal{L}^k which embeds a modal operator K to represent an agent's knowledge about access control policies.

2.1 Syntax of \mathcal{L}^k

Language \mathcal{L}^k includes the following disjoint sorts for *subject*, *group-subject*, *access-right*, *group-access-right*, *object*, *group-object* together with predicate symbols *holds*, \in , \subseteq and logic connectives \wedge and \neg .

The six disjoint sorts and the predicate symbols \in and \subseteq are defined as follows:

1. Sort *subject*: with subject constants S, S_1, S_2, \dots , and subject variables s, s_1, s_2, \dots .
2. Sort *group-subject*: with group subject constants G, G_1, G_2, \dots , and group subject variables g, g_1, g_2, \dots .
3. Sort *access-right*: with access right constants A, A_1, A_2, \dots , and access right variables a, a_1, a_2, \dots .
4. Sort *group-access-right*: with group access right constants GA, GA_1, GA_2, \dots , and group access right variables ga, ga_1, ga_2, \dots .
5. Sort *object*: with object constants O, O_1, O_2, \dots , and object variables o, o_1, o_2, \dots .
6. Sort *group-object*: with group object constants GO, GO_1, GO_2, \dots , and group object variables go, go_1, go_2, \dots .
7. A modal operator K to represent what an agent *knows* to be true.
8. A ternary predicate symbol *holds* which takes arguments as *subject* or *group-subject*, *access-right* or *group-access-right* and *object* or *group-object* respectively, which is used to represent particular access rights with the associated subjects and objects.
9. A binary predicate symbol \in which takes arguments as *subject* and *group-subject* or *access-right* and *group-access-right* or *object* and *group-object* respectively.
10. A binary predicate symbol \subseteq whose both arguments are *group-subjects*, *group-access-rights* or *group-objects*.

In \mathcal{L}^k , we define a *fact* f to be an atomic formula or its negation. A *ground fact* is a fact without variable occurrence. We view $\neg\neg f$ as f . A *fact expression* ϕ of \mathcal{L}^k is defined as follows: (i) each fact ϕ is a fact expression; (ii) if ϕ and ψ are fact expressions, then $\phi \wedge \psi$ and $\phi \vee \psi$ are also fact expressions. A *ground fact expression* is a fact expression without variable occurrence. A ground fact expression is called a *ground instance* of a fact expression if this ground fact expression is obtained from the fact expression by replacing each of its variable occurrence with the same sort constant. A fact expression ϕ is called *conjunctive* (or *disjunctive*) if it is of the form $\phi_1 \wedge \cdots \wedge \phi_n$ (or $\phi_1 \vee \cdots \vee \phi_n$ respectively), where each ϕ_i is a fact.

Now we are ready to formally define the propositions in \mathcal{L}^k . An *initial proposition* in \mathcal{L}^k is defined as

$$\text{initially } \phi \tag{1}$$

where ϕ is either a conjunctive or disjunctive fact expression. That is, ϕ is of the form $\phi_1 \wedge \cdots \wedge \phi_n$ or $\phi_1 \vee \cdots \vee \phi_n$, where each ϕ_i is a fact.

An *objective proposition* is an expression of the form

$$\phi \text{ if } \psi \text{ with absence } \gamma \tag{2}$$

where ϕ is either a conjunctive or disjunctive fact expression, ψ and γ are two conjunctive fact expressions.

A *subjective proposition* is an expression of the form

$$\phi \text{ if } \psi \text{ with absence } \gamma \text{ knowing } \beta, \tag{3}$$

or

$$\phi \text{ if } \psi \text{ with absence } \gamma \text{ not knowing } \beta, \tag{4}$$

where ϕ is a conjunctive or disjunctive fact expression, and ψ , γ and β are conjunctive fact expressions.

A proposition is called a *ground proposition* if it does not contain variables. A *policy domain description* D in \mathcal{L}^k is a finite set of initial propositions, objective propositions and subjective propositions.

2.2 Representing Complex Access Control Scenarios Using \mathcal{L}^K

In the following, we describe a few complex security scenarios using language \mathcal{L}^K , and demonstrate that \mathcal{L}^K is an expressive language to represent incomplete information, default information, and agents' knowledge in relation to various access control situations.

Example 1. Consider a scenario where a department has some classified files related to different projects. The security rules are: the directors of the department can access all these classified files, all other staff can only access the files related to the project he is currently assigned. Assume that Staff, Director, ProjectA represent the group of all staff of the department, the group of directors and the group of staff working for project A respectively. If we know that Alice is

a director, Bob is a staff working for project A and Carl is a staff working for other project. FileA is a classified document of Project A.

A domain description D specifies this scenario consists of the following propositions:

initially $holds(Alice, Access, FileA)$,
initially $holds(Bob, Access, FileA)$,
initially $\neg holds(Carl, Access, FileA)$,
 $holds(x, Access, FileA)$ **if** $x \in Staff \wedge x \in Director$,
 $holds(x, Access, FileA)$ **if** $x \in Staff \wedge x \in ProjectA$,
 $\neg holds(x, Access, FileA)$ **if** $x \in Staff \wedge \neg(x \in Director) \wedge \neg(x \in ProjectA)$.

Example 2. Consider a domain description D consists of the following propositions:

initially $holds(S, Own, O)$,
 $holds(S, Write, O)$ **if** $holds(S, Own, O)$
with absence $\neg holds(S, Write, O)$,
 $\neg holds(S, Own, O)$ **if** $\neg holds(S, Read, O)$.

This domain description expresses the following policies: initially subject S owns object O . If there is no evidence that S cannot write on O is absent from the domain, then S has write right on O , and S will no longer owns O if somehow S cannot read O anymore. Here **with absence** $\neg holds(S, Write, O)$ represents a default information. As long as there is no clear information indicating $\neg holds(S, Write, O)$, it would be assumed that S can write O .

3 Semantics of \mathcal{L}^k

Given a domain description D , we will translate it into an epistemic logic program $\Pi(D)$, then the semantics of D will be defined based on the *world view semantics* of program $\Pi(D)$. In the following, we first introduce epistemic logic programs, and then define the semantics of \mathcal{L}^k . Due to space limitation, we will skip the translation details and just focus on the epistemic logic programs.

In this section, we present a general overview on epistemic logic programs. Gelfond extended the syntax and semantics of disjunctive logic programs to allow the correct representation of incomplete information in the presence of multiple extensions [7]. In epistemic logic programs, the language of (disjunctive) extended logic programs is expanded with two modal operators K and M . KF is read as “ F is known to be true” and MF is read as “ F may be believed to be true”. In this paper we will consider propositional epistemic logic programs where rules containing variables are viewed as the set of all ground rules by replacing these variables with all constants occurring in the language. The semantics for epistemic logic programs is defined by the pair (\mathcal{A}, W) , where \mathcal{A} is a collection of sets of ground literals which is also simply called is a collection of *belief sets*, and W is a set in \mathcal{A} called the agent’s *working set of beliefs*. The truth of a

formula F in (\mathcal{A}, W) is denoted by $(\mathcal{A}, W) \models F$ and the falsity is denoted by $(\mathcal{A}, W) = \!| F$. They are defined as follows.

- $(\mathcal{A}, W) \models p$ iff $p \in W$ where p is a propositional atom.
- $(\mathcal{A}, W) \models KF$ iff $(\mathcal{A}, W_i) \models F$ for all $W_i \in A$.
- $(\mathcal{A}, W) \models MF$ iff $(\mathcal{A}, W_i) \models F$ for some $W_i \in A$.
- $(\mathcal{A}, W) \models F \wedge G$ iff $(\mathcal{A}, W) \models F$ and $(\mathcal{A}, W) \models G$.
- $(\mathcal{A}, W) \models F \text{ or } G$ iff $(\mathcal{A}, W) \models \neg(\neg F \wedge \neg G)$.
- $(\mathcal{A}, W) \models \neg F$ iff $(\mathcal{A}, W) = \!| F$.
- $(\mathcal{A}, W) = \!| F$ iff $\neg F \in W$ where F is a ground atom.
- $(\mathcal{A}, W) = \!| KF$ iff $(\mathcal{A}, W) \not\models KF$. $(\mathcal{A}, W) = \!| MF$ iff $(\mathcal{A}, W) \not\models MF$.
- $(\mathcal{A}, W) = \!| F \wedge G$ iff $(\mathcal{A}, W) = \!| F$ or $(\mathcal{A}, W) = \!| G$.
- $(\mathcal{A}, W) = \!| F \text{ or } G$ iff $(\mathcal{A}, W) = \!| F$ and $(\mathcal{A}, W) = \!| G$.

It is worth mentioning that since belief set W allows both positive and negative propositional atoms, in Gelfond's semantics, $(\mathcal{A}, W) = \!| \varphi$ is not equivalent to $(\mathcal{A}, W) \not\models \varphi$ in general. For instance, $(\{\{a, b\}\}, \{a, b\}) \not\models c$, but we do not have $(\{\{a, b\}\}, \{a, b\}) = \!| c$ (i.e. $(\{\{a, b\}\}, \{a, b\}) \models \neg c$). Consequently, here K and M are *not* dual modal operators here. Consider $\mathcal{A} = \{\{a, b\}, \{a, b, \neg c\}\}$. Clearly we have $\mathcal{A} \models \neg K\neg c$. But having $\mathcal{A} \models Mc$ seems to be wrong.

If a formula G is of the form KF , $\neg KF$, MF or $\neg MF$ (where F is a propositional formula), then its truth value in (\mathcal{A}, W) will not depend on W . In this case we call G a *subjective formula*. If F is a propositional literal, then we call KF , $\neg KF$, MF , and $\neg MF$ *subjective literals*. On the other hand, if G does not contain K or M , then its truth value in (\mathcal{A}, W) will only depend on W and we call G an *objective formula* or objective literal if G is a propositional literal. In the case that G is subjective, we simply write $\mathcal{A} \models G$ instead of $(\mathcal{A}, W) \models G$, and $W \models G$ instead of $(\mathcal{A}, W) \models G$ in the case that G is objective. In general, we simply write $A \models G$ if for each $W \in A$, we have $(\mathcal{A}, W) \models G$. each A

An *epistemic logic program* Π is a finite set of rules of the form:

$$F \leftarrow G_1, \dots, G_m, \text{not } G_{m+1}, \dots, \text{not } G_n. \quad (5)$$

In (5), $m, n \geq 0$, F is of the form $F_1 \text{ or } \dots \text{ or } F_k$ ($k \geq 1$) and F_1, \dots, F_k are objective literals, G_1, \dots, G_m are objective or subjective literals, and G_{m+1}, \dots, G_n are objective literals. For an epistemic logic program \mathcal{P} , its semantics is given by its *world view* which is defined in the following steps:

Step 1. Let Π be an epistemic logic program not containing modal operators \overline{K} and \overline{M} and negation as failure *not*. A set W of ground literals is called a *belief set of Π* iff W is a minimal set of satisfying conditions: (i) for each rule $F \leftarrow G_1, \dots, G_m$ from Π such that $W \models G_1 \wedge \dots \wedge G_m$ we have $W \models F$; and (ii) if W contains a pair of complementary literals then $W = \text{Lit}$, i.e. W is an inconsistent belief set.

Step 2. Let Π be an epistemic logic program not containing modal operators \overline{K} and \overline{M} and W be a set of ground literals in the language of Π . By Π_W we denote the result of (i) removing from Π all the rules containing formulas of the

form *not G* such that $W \models G$ and (ii) removing from the rules in Π all other occurrences of formulas of the form *notG*.

Step 3. Finally, let Π be an arbitrary epistemic logic program and \mathcal{A} a collection of sets of ground literals in its language. By $\Pi_{\mathcal{A}}$ we denote the epistemic logic program obtained from Π by (i) removing from Π all rules containing formulas of the form G such that G is subjective and $\mathcal{A} \not\models G$, and (ii) removing from rules in Π all other occurrences of subjective formulas. Now we define that a collection \mathcal{A} of sets of ground literals is a *world view* of Π if \mathcal{A} is the collection of all belief sets of $\Pi_{\mathcal{A}}$.

4 A Case Study: Reasoning about Knowledge in Access Control

So far we have specified the semantics of our access control policy language \mathcal{L}^K which can represent an agent's knowledge about policies. In this section, we will demonstrate a non-trivial case study, from which we show that our approach for knowledge based access control can overcome some difficulties in the reasoning about access control when incomplete information is involved.

Consider a scenarios that a company's IT administrator needs to specify various policies about staff's authorization to access the company's business profile. Basically, if a staff is a project manager, then he/she should be allowed to access the company's business profile; if a staff is a business analyser *and* a team leader for some project, then he/she should be also allowed to access the company business profile. On the other hand, if a staff is neither a project manager nor a team leader, then he/she should not be allowed to access the company's business profile. Finally, if there is no information showing whether a staff can access the company business profile or not, the administrator should mark a "waiting for approval" status for this staff and only allows him/her to access the previous years company business profile.

By using our language \mathcal{L}^K , we can encode the above policies into the following propositions:

$$\text{holds}(x, \text{Read}, \text{Profile}) \text{ if } \text{projectManager}(x), \quad (6)$$

$$\text{holds}(x, \text{Read}, \text{Profile}) \text{ if } \text{businessAnalyser}(x) \wedge \text{teamLeader}(x), \quad (7)$$

$$\neg \text{holds}(x, \text{Read}, \text{Profile}) \text{ if } \neg \text{projectManager}(x) \wedge \neg \text{teamLeader}(x), \quad (8)$$

$$\text{waitingApproval}(x) \text{ if with absence}$$

$$\text{holds}(x, \text{Read}, \text{Profile}) \wedge \neg \text{holds}(x, \text{Read}, \text{Profile}), \quad (9)$$

$$\text{holds}(x, \text{Read}, \text{oldProfile}) \text{ if } \text{waitingApproval}(x), \quad (10)$$

Note that in the above propositions, no subjective proposition is used. Now suppose currently the IT administrator only has incomplete information that Bob is either a project manager or a team leader, but does not know exactly what role he holds, as stated by the following initial proposition:

$$\text{initially } \text{projectManager}(\text{Bob}) \vee \text{teamLeader}(\text{Bob}) \quad (11)$$

Let D be the domain description consisting of propositions (6) - (III). Then we want to know Bob is allowed to access the current business profile or the old business profile. To answer such query, we first translate D into the following epistemic logic program $\Pi(D)$:

$$\begin{aligned}
r_1 &: \text{holds}(x, \text{Read}, \text{Profile}) \leftarrow \text{projectManager}(x), \\
r_2 &: \text{holds}(x, \text{Read}, \text{Profile}) \leftarrow \text{businessAnalyser}(x), \text{teamLeader}(x), \\
r_3 &: \neg \text{holds}(x, \text{Read}, \text{Profile}) \leftarrow \neg \text{projectManager}(x), \neg \text{teamLeader}(x), \\
r_4 &: \text{waitingApproval}(x) \leftarrow \text{not holds}(x, \text{Read}, \text{Profile}), \\
&\quad \text{not } \neg \text{holds}(x, \text{Read}, \text{Profile}), \\
r_5 &: \text{holds}(x, \text{Read}, \text{oldProfile}) \leftarrow \text{waitingApproval}(x), \\
r_6 &: \text{projectManager}(\text{Bob}) \vee \text{teamLeader}(\text{Bob}) \leftarrow.
\end{aligned}$$

Quite easily to see that $\Pi(D)$ has a unique world view

$$\begin{aligned}
A = \{ & \{ \text{holds}(\text{Bob}, \text{Read}, \text{Profile}), \text{projectManager}(\text{Bob}) \}, \\
& \{ \text{teamLeader}(\text{Bob}), \text{waitingApproval}(\text{Bob}), \\
& \text{holds}(\text{Bob}, \text{Read}, \text{oldProfile}) \} \},
\end{aligned}$$

from which we have $\Pi(D) \not\models \text{holds}(x, \text{Read}, \text{Profile})$ as well as $\Pi(D) \not\models \text{holds}(x, \text{Read}, \text{oldProfile})$, that means that the administrator cannot decide whether he should allow Bob to access either the current profile or the old profile.

This is a quite weak conclusion from D because intuitively since there is no explicit information about whether Bob can access the current profile or not, we should conclude that the administrator needs to mark Bob as in a "waiting for approval" status and hence allows him to access old profile at this stage. Unfortunately, such intuition has not been encoded into the corresponding description D .

By using a proper subjective proposition, this difficulty can be overcome in \mathcal{L}^K . We simply replace proposition (9) by the following subjective proposition:

$$\begin{aligned}
& \text{waitingApproval}(x) \text{ if not knowing} \\
& \text{holds}(x, \text{Read}, \text{Profile}) \wedge \neg \text{holds}(x, \text{Read}, \text{Profile}), \tag{12}
\end{aligned}$$

and we specify a new domain description D' consisting of propositions (6) - (8), (10) - (12). Then we have $\Pi(D')$ consists of rules of r_1, r_2, r_3, r_5, r_6 together with the following $r_{4'}$:

$$\begin{aligned}
r_{4'} &: \text{waitingApproval}(x) \leftarrow \neg K \text{holds}(x, \text{Read}, \text{Profile}), \\
&\quad \neg K \neg \text{holds}(x, \text{Read}, \text{Profile}).
\end{aligned}$$

$\Pi(D')$ has a unique world view

$$\begin{aligned}
A' = \{ & \{ \text{holds}(\text{Bob}, \text{Read}, \text{Profile}), \text{projectManager}(\text{Bob}), \\
& \text{waitingApproval}(\text{Bob}), \text{holds}(\text{Bob}, \text{Read}, \text{oldProfile}) \}, \\
& \{ \text{teamLeader}(\text{Bob}), \text{waitingApproval}(\text{Bob}), \\
& \text{holds}(\text{Bob}, \text{Read}, \text{oldProfile}) \} \},
\end{aligned}$$

from which we can derive $\Pi(D') \models \text{holds}(\text{Bob}, \text{Read}, \text{oldProfile})$, that is, Bob is allowed to access the old profile.

This case study has showed an important application of knowledge based propositions in access control specifications using our language \mathcal{L}^K while most of current access control formal languages cannot deal with properly.

5 Conclusions

In this paper, we proposed a language \mathcal{L}^k to specify security policies by an authorization domain with incomplete information. We introduced modal logic for the specification of incomplete domain and applied epistemic logic program for the reasoning. We also showed that our approach has an expressive power to describe a variety of complex security scenarios. At this stage, our work is limited to a single agent with a single authorization domain managing a centralized information system. For a multi-agent, distributed system environment, it needs a different approach to handle multi-domain specification and reasoning, authorization delegation and delegation depth etc. It is our future work to implement the approach presented in this work and to consider multi-agent presentation.

References

1. Bai, Y., Varadharajan, V.: On transformation of authorization policies. *Data and Knowledge Engineering* 45(3), 333–357 (2003)
2. Bertino, E., Catania, B., Ferrari, E., Perlasca, P.: A logical framework for reasoning about access control models. *ACM Transactions on Information and System Security* 6(1), 71–127 (2003)
3. Chomicki, J., Lobo, J., Naqvi, S.: A logical programming approach to conflict resolution in policy management. In: *Proceedings of International Conference on Principles of Knowledge Representation and Reasoning*, pp. 121–132 (2000)
4. Crampton, J., Khambhammettu, H.: Delegation in role-based access control. *International Journal of Information Security* 7, 123–136 (2008)
5. Denning, D.E.: A lattice model of secure information flow. *Communication of ACM* 19, 236–243 (1976)
6. Fagin, R., Halpern, J.Y., Moses, Y., Vardi, M.Y.: *Reasoning about knowledge*. MIT Press, Cambridge (1995)
7. Gelfond, M.: Logic programming and reasoning with incomplete information. *Annals of Mathematics and Artificial Intelligence* 12, 98–116 (1994)
8. Meadows, C.: Policies for Dynamic Upgrading. In: *Database Security, IV: Status and Prospects*, pp. 241–250 (1991)
9. Murray, T., Grove, D.: Non-delegatable authorities in capability systems. *Journal of Computer Security* 16, 743–759 (2008)
10. Zhang, Y.: Epistemic reasoning in logic programs. In: *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI 2007)*, pp. 647–652 (2007)
11. Zhou, J., Alves-Foss, J.: Security policy refinement and enforcement for the design of multi-level secure systems. *Journal of Computer Security* 16, 107–131 (2008)

Multi Criteria Decision Making in Fuzzy Description Logics: A First Step

Umberto Straccia

Istituto di Scienza e Tecnologie dell'Informazione (ISTI - CNR), Pisa, Italy
straccia@isti.cnr.it

Abstract. Fuzzy Description Logics are logics which allow to deal with structured knowledge affected by vagueness. Although a relatively important amount of work has been carried out in the last years, fuzzy DLs are open to be extended with several features worked out in other fields. In this work, we start addressing the problem of incorporating Multi-Criteria Decision Making (MCDM) into fuzzy Description Logics and, thus, start an investigation about offering the possibility of a fuzzy ontology assisted approach to decision making.

1 Introduction

Description Logics (DLs) [1] play a key role in the design of *Ontologies*. An ontology consists of a hierarchical description of important concepts in a particular domain, along with the description of the properties (of the instances) of each concept. DLs play a particular role in this context as they are essentially the theoretical counterpart of the *Web Ontology Language OWL DL*, a state of the art language to specify ontologies.

It is well-known that “classical” ontology languages are not appropriate to deal with *fuzzy knowledge*, which is inherent to several real world domains [10,12]. Fuzzy ontologies emerge as useful in several applications, such as (multimedia) information retrieval, image interpretation, ontology mapping, matchmaking and the Semantic Web [8]. So far, several fuzzy extensions of DLs can be found in the literature (see the survey in [8]) and some fuzzy DL reasoners have been implemented, such as FUZZYDL [3], DELOREAN [2] or FIRE [9].

In this work, we start investigating about using fuzzy DLs as a fuzzy ontology support for *Multi-Criteria Decision Making* (MCDM) [13], which is among one of the most well known branches of decision making. Roughly, MCDM is the study of identifying and choosing alternatives based on the values and preferences of the decision maker. Making a decision implies that there are alternative choices to be considered and to choose the one that best fits with our goals, objectives, desires, values, and so on. Our work should be understood as an attempt in using fuzzy DLs and, thus, fuzzy ontologies, for knowledge assisted decision making. While there is a large literature on fuzzy MCDM [6] and fuzzy DLs [8], to the best of our knowledge, this is the first time such a combination is addressed.

We proceed as follows. Section 2 (resp. Section 3) will provide the basic concepts related to mathematical fuzzy logic (resp. MCDM) we will rely on, Section 4

specifies a minimal fuzzy DL to deal with MCDM and illustrates some examples. Section 5 concludes and describes some future work.

2 Fuzzy Sets and Mathematical Fuzzy Logic Basics

In *Mathematical Fuzzy Logic* [5], the convention prescribing that a statement is either true or false is changed and is a matter of degree measured on an ordered scale \mathcal{S} that is no longer $\{0, 1\}$, but usually the unit interval $[0, 1]$. This degree of fit is called *degree of truth* of the statement ϕ in the interpretation \mathcal{I} . In this section, for illustrative purposes, *fuzzy statements* have the form $\phi \geq l$ or $\phi \leq u$, where $l, u \in [0, 1]$ (see, e.g. [5]) and ϕ is a statement. Fuzzy statements encode that the degree of truth of ϕ is *at least equal to* l resp. *at most equal to* u . A *fuzzy interpretation* \mathcal{I} maps each basic statement p_i into $[0, 1]$ and is then extended inductively to all statements: $\mathcal{I}(\phi \wedge \psi) = \mathcal{I}(\phi) \otimes \mathcal{I}(\psi)$, $\mathcal{I}(\phi \vee \psi) = \mathcal{I}(\phi) \oplus \mathcal{I}(\psi)$, $\mathcal{I}(\phi \rightarrow \psi) = \mathcal{I}(\phi) \Rightarrow \mathcal{I}(\psi)$, $\mathcal{I}(\neg\phi) = \ominus \mathcal{I}(\phi)$, $\mathcal{I}(\exists x.\phi(x)) = \sup_{a \in \Delta^{\mathcal{I}}} \mathcal{I}(\phi(a))$, $\mathcal{I}(\forall x.\phi(x)) = \inf_{a \in \Delta^{\mathcal{I}}} \mathcal{I}(\phi(a))$, where $\Delta^{\mathcal{I}}$ is the domain of \mathcal{I} , and \otimes , \oplus , \Rightarrow , and \ominus are *t-norms*, *t-conorms*, *implication functions*, and *negation functions*, respectively, which extend the classical Boolean conjunction, disjunction, implication, and negation, respectively, to the fuzzy case. One usually distinguishes three different logics (see below), namely Lukasiewicz, Gödel, and Product logic [5]. Zadeh “logic”, namely $a \otimes b = \min(a, b)$, $a \oplus b = \max(a, b)$, $\ominus a = 1 - a$ and $a \Rightarrow b = \max(1 - a, b)$ is entailed by Lukasiewicz logic.

	Lukasiewicz Logic	Gödel Logic	Product Logic		Lukasiewicz Logic	Gödel Logic	Product Logic
$a \otimes b$	$\max(a + b - 1, 0)$	$\min(a, b)$	$a \cdot b$	$a \Rightarrow b$	$\min(1 - a + b, 1)$	$\begin{cases} 1 & \text{if } a \leq b \\ b & \text{otherwise} \end{cases}$	$\min(1, b/a)$
$a \oplus b$	$\min(a + b, 1)$	$\max(a, b)$	$a + b - a \cdot b$	$\ominus a$	$1 - a$	$\begin{cases} 1 & \text{if } a = 0 \\ 0 & \text{otherwise} \end{cases}$	$\begin{cases} 1 & \text{if } a = 0 \\ 0 & \text{otherwise} \end{cases}$

In *fuzzy set theory* [7], a *fuzzy set* R over a countable crisp set X is a function $R: X \rightarrow [0, 1]$. A (binary) *fuzzy relation* R over two countable crisp sets X and Y is a function $R: X \times Y \rightarrow [0, 1]$. We say that R is *functional* iff R is a partial function $R: X \times Y \rightarrow \{0, 1\}$ such that for each $x \in X$ there is unique $y \in Y$ where $R(x, y)$ is defined. The *degree of subsumption* between two fuzzy sets A and B is defined as $\inf_{x \in X} A(x) \Rightarrow B(x)$ and may be seen as the degree of the FOL formula $\forall x.A(x) \rightarrow B(x)$, while the *degree of overlap* between two fuzzy sets A and B is defined as $\sup_{x \in X} A(x) \wedge B(x)$ and may be seen as the degree of the FOL formula $\exists x.A(x) \wedge B(x)$.

The notions of satisfiability and logical consequence are defined in the standard way. A fuzzy interpretation \mathcal{I} *satisfies* a fuzzy statement $\phi \geq l$ (resp., $\phi \leq u$) or \mathcal{I} is a *model* of $\phi \geq l$ (resp., $\phi \leq u$), denoted $\mathcal{I} \models \phi \geq l$ (resp., $\mathcal{I} \models \phi \leq u$), iff $\mathcal{I}(\phi) \geq l$ (resp., $\mathcal{I}(\phi) \leq u$). Furthermore, $\phi \geq l$ is a *tight logical consequence* of a set of fuzzy statements \mathcal{K} iff l is the infimum of $\mathcal{I}(\phi)$ subject to all models \mathcal{I} of \mathcal{K} . The latter value is equivalent to $l = \sup \{r \mid \mathcal{K} \models \phi \geq r\}$, it is called *Best Entailment Degree* (BED), and is denoted $bed(\mathcal{K}, \phi)$, while the *Best Satisfiability Degree* (BSD), denoted as $bsd(\mathcal{K}, \phi)$, is defined as $\sup_{\mathcal{I} \models \mathcal{K}} \mathcal{I}(\phi)$.

3 MCDM Basics

The area of MCDM is quite vast and we cannot address all the addressed issues here. We will focus on the basic notions that are of importance in MCDM and a simple MCDM method to be used here.

Usually, *alternatives* represent different choices of action available to the decision maker and is assumed to be finite in our case. The *decision criteria* represent the different dimensions from which the alternatives can be viewed (a decision criteria is also referred to as *goals* or *attributes*). Most of the MCDM methods require the criteria be assigned *decision weights* of importance. Usually, these weights are normalized to add up to one.

A *MCDM problem* of m criteria and n alternatives is informally as follows: let $\mathbf{A} = \{A_1, \dots, A_n\}$ be a set of n decision alternatives and let $\mathbf{C} = \{C_1, \dots, C_m\}$ be a set of m criteria according to which the desirability of an action is judged. Determine the optimal alternative A^* with the highest degree of desirability.

A standard feature of MCDM methods is that a MCDM problem can be expressed by means of a *decision matrix*, as shown below.

		Criteria				
		w_1	w_2	\cdot	\cdot	w_m
Alternatives		C_1	C_2	\cdot	\cdot	C_m
x_1	A_1	a_{11}	a_{12}	\cdot	\cdot	a_{1m}
x_2	A_2	a_{21}	a_{22}	\cdot	\cdot	a_{2m}
\cdot	\cdot	\cdot	\cdot	\cdot	\cdot	\cdot
\cdot	\cdot	\cdot	\cdot	\cdot	\cdot	\cdot
x_n	A_n	a_{n1}	a_{n2}	\cdot	\cdot	a_{nm}

(1)

In the matrix each column belongs to a criterion C_j and each row describes the performance of an alternative A_i . The score a_{ij} describes the performance of alternative A_i against criterion C_j . The weights w_1, \dots, w_m are assigned to the criteria. Weight w_j reflects the relative importance of criteria C_j to the decision, and is assumed to be positive and normalized, i.e. $1 = \sum_{j=1}^m w_j$. The weights of the criteria are usually determined on subjective basis and may also be seen as a kind of profit of the criteria. They represent the opinion of a single decision maker or synthesize the opinions of a group of experts. Not surprisingly, there is a large literature on methods to assign weights (see, e.g. [13]). For illustrative purpose, we illustrate here a method based on pairwise comparison of the criteria to determine the weights. It consists in comparing elements (criteria) X_i with X_j ($1 \leq i, j \leq k$) and judge how much they contribute to the overall objective. The judgment consists in assigning a number $w_{ij} \in [1, 9]$, called *Intensity of Importance*, selected according the following table

Intensity	Definition	Explanation
1	Equal Importance	Two elements contribute equally to the objective
3	Moderate Importance	Experience and judgement slightly favor one element over the other
5	Strong Importance	Experience and judgement strongly favor one element over the other
7	Very Strong Importance	One element is favored very strongly over another, its dominance is demonstrated in practice
9	Extreme Importance	The evidence favoring one element over another is of the highest possible order of information

Intermediate values can be used. The weight w_i of element X_i may be obtained as

$$\bar{w}_i = \left(\prod_{j=1}^k w_{ij} \right)^{1/k}, \quad \bar{w} = \sum_{i=1}^k \bar{w}_i, \quad \text{and then } w_i = \bar{w}_i / \bar{w}.$$

On more on alternatives to determine the data of a decision matrix and their impact, see e.g., [13].

The values x_1, \dots, x_n associated with the alternatives in the decision matrix will be used to denote the final ranking values of the alternatives. Usually, higher ranking value means a better performance of the alternative, so the alternative with the highest ranking value is the best of the alternatives. MCDM techniques can partially or completely rank the alternatives: a single most preferred alternative can be identified or a short list of a limited number of alternatives can be selected for subsequent detailed appraisal. Again, there are many alternative methods to compute the final ranking values from the decision matrix. For illustrative purposes, we present the so-called *Weighted Sum Method* (WSM), which is among the simplest methods in MCDM, but has the advantage to be easy embedded within fuzzy DLs. Formally, let

$$x_i = \sum_{j=1}^m a_{ij} w_j \quad \text{for } i = 1, 2, \dots, m. \quad (2)$$

where x_i is the the *final ranking value* of alternative A_i . The *ranking of the alternatives* is obtained by ordering the alternatives in descending order with respect to the final ranking value and the *optimal alternative* A^* is the one that maximizes the final ranking value, i.e.

$$A^* = \arg \max_{A_i} x_i.$$

We conclude this section by pointing out that in *fuzzy* MCDM, a principal difference to classical MCDM is due to the fact that weights w_i and performance factors a_{ij} are so-called *fuzzy numbers* [7]. A fuzzy number \tilde{n} is a fuzzy set over reals with triangular membership function $\text{tri}(a, b, c)$ and is intended being an approximation of the number b . Any real value n is seen as the fuzzy number $\text{tri}(n, n, n)$. The arithmetic operators $+$, $-$, \cdot and \div are extended to fuzzy numbers by applying them to the arguments, i.e. for fuzzy numbers $\tilde{n}_1 = \text{tri}(a_1, b_1, c_1)$ and $\tilde{n}_2 = \text{tri}(a_2, b_2, c_2)$, for operator $*$ $\in \{+, \cdot\}$, $\tilde{n}_1 * \tilde{n}_2 = \text{tri}(a_1 * a_2, b_1 * b_2, c_1 * c_2)$, while for $*$ $\in \{-, \div\}$, $\tilde{n}_1 * \tilde{n}_2 = \text{tri}(a_1 * c_2, b_1 * b_2, c_1 * a_2)$. The final rank value is computed as in Eq. [2].

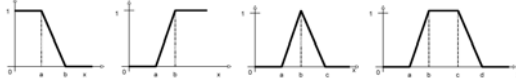
$$\tilde{x}_i = \sum_{j=1}^m \tilde{a}_{ij} \cdot \tilde{w}_j \quad \text{for } i = 1, 2, \dots, m. \quad (3)$$

As now \tilde{x}_i is a fuzzy number, one may apply some defuzzification method to it and compare fuzzy numbers based on these values, or one may use some fuzzy number comparison operator to determine the optimal solution (see, e.g. [13]).

4 Towards MCDM within Fuzzy Description Logics

The aim of this section is to show that current fuzzy DLs can be used to deal with (fuzzy) knowledge assisted MCDM (though, the MCDM method needs to be simple so far). For illustrative purposes and for reasons of space, we will just consider a minimal fuzzy DL to deal with the WSM in MCDM.

Syntax. We will present $\mathcal{ALCF}(D)$, which is the basic DL \mathcal{ALC} extended with functional roles (letter \mathcal{F}) and concrete domains $\boxed{\text{II}}$ (letter D).



In general, a *fuzzy concrete domain* (or simply *fuzzy domain*) $\boxed{\text{II}}$ is a pair $\langle \Delta_D, \Phi_D \rangle$, where Δ_D is an interpretation domain and Φ_D is the set of *fuzzy domain predicates* d with a predefined arity n and an interpretation $d^D : \Delta_D^n \rightarrow [0, 1]$, which is a n -ary fuzzy relation over Δ_D . In our specific spatial fuzzy DL, we assume that predicates are unary and Δ_D are non-negative real numbers.

Now, consider pairwise disjoint alphabets of *concepts names* (denoted A), *abstract roles names* (denoted R) and *concrete roles names* (denoted T). Within the alphabet of abstract and concrete roles, we have distinguished subsets of *abstract functional roles names* (denoted f) and *concrete functional roles names* (denoted t), respectively. We call functional roles also *features*. From a First-Order Logic point of view, concepts may be seen as a formulae with one free variable (and, thus, may be seen as class descriptors), while roles as binary predicates (and, thus, may be used to describe properties of a class). *Concepts* (denoted C or D) of the language can be built inductively from atomic concepts (A), top concept \top , bottom concept \perp , abstract roles (R), concrete roles (T) as follows. The syntax of fuzzy concepts (denoted C, D) is as follows:

$$C, D := \top \mid \perp \mid A \mid C \sqcap D \mid C \sqcup D \mid \neg C \mid \forall R.C \mid \exists R.C$$

Now, the fuzzy DL is extended as follows $\boxed{\text{3}}$:

$$\begin{aligned} C, D &:= \forall T.d \mid \exists T.d \mid w_1 C_1 + \dots + w_k C_k \\ d &:= ls(a, b) \mid rs(a, b) \mid tri(a, b, c) \mid trap(a, b, c, d) \end{aligned}$$

where val is an integer, a real or a string depending on the range of the concrete feature t , $w_i \in [0, 1]_D$, $\sum_{i=1}^k w_i \leq 1$ and C_i are concepts¹. E.g., the expression $Human \sqcap (\leq hasAge \ 18)$ will denote the set of humans, which have an age less or equal than 18, while $Human \sqcap \exists hasAge.ls(10, 30)$ will denote the set of young humans (their age is $ls(10, 30)$).

A *Fuzzy Knowledge Base* (or *fuzzy Ontology*) consists of a finite set of *fuzzy General Concept Inclusions* (*fuzzy GCIs*), which are expressions of the form $\langle C \sqsubseteq D, n \rangle$ (with informal meaning, the degree of subsumption between concept C and D is not less than n). In FOL, $\langle C \sqsubseteq D, n \rangle$ may be seen as a fuzzy statement of the form $(\forall x.C(x) \rightarrow D(x)) \geq n$ and amounts of asserting that the degree of

¹ In $\boxed{\text{3}}$ we assume $\sum_{i=1}^k w_i = 1$ instead, however, this modification is harmless.

subsumption among C and D is at least n . We will use $C = D$ as a shorthand for $\langle C \sqsubseteq D, 1 \rangle$ and $\langle D \sqsubseteq C, 1 \rangle$.

Semantics. From a semantics point of view, a *fuzzy interpretation* $\mathcal{I} = (\Delta^{\mathcal{I}}, \cdot^{\mathcal{I}})$ relative to the fuzzy concrete domain $\langle \Delta_D, \Phi_D \rangle$, consists of a nonempty set $\Delta^{\mathcal{I}}$ (the *domain*), disjoint from Δ_D , and of a *fuzzy interpretation function* $\cdot^{\mathcal{I}}$ that coincides with \cdot_D on every fuzzy concrete predicate, and it assigns: (i) to each abstract concept C a function $C^{\mathcal{I}}: \Delta^{\mathcal{I}} \rightarrow [0, 1]$; (ii) to each abstract role R a function $R^{\mathcal{I}}: \Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}} \rightarrow [0, 1]$; (iii) to each abstract feature r a partial function $r^{\mathcal{I}}: \Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}} \rightarrow \{0, 1\}$ such that for all $u \in \Delta^{\mathcal{I}}$ there is a unique $w \in \Delta^{\mathcal{I}}$ on which $r^{\mathcal{I}}(u, w)$ is defined; (iv) to each concrete role T a function $T^{\mathcal{I}}: \Delta^{\mathcal{I}} \times \Delta_D \rightarrow [0, 1]$; (v) to each concrete feature t a partial function $t^{\mathcal{I}}: \Delta^{\mathcal{I}} \times \Delta_D \rightarrow \{0, 1\}$ such that for all $u \in \Delta^{\mathcal{I}}$ there is a unique $r \in \Delta_D$ on which $t^{\mathcal{I}}(u, r)$ is defined.

Given arbitrary t-norm \otimes , t-conorm \oplus , negation function \ominus and implication function \Rightarrow , the fuzzy interpretation function is extended to *complex concepts* and *fuzzy axioms* as below:

$$\begin{aligned}
(\top)^{\mathcal{I}}(x) &= 1 & (\exists R.C)^{\mathcal{I}}(x) &= \sup_{y \in \Delta^{\mathcal{I}}} \{R^{\mathcal{I}}(x, y) \otimes C^{\mathcal{I}}(y)\} \\
(\perp)^{\mathcal{I}}(x) &= 0 & (\forall T.d)^{\mathcal{I}}(x) &= \inf_{r \in \Delta_D} \{T^{\mathcal{I}}(x, r) \Rightarrow d^{\mathcal{I}}(r)\} \\
(A)^{\mathcal{I}}(x) &= A^{\mathcal{I}}(x) & (\exists T.d)^{\mathcal{I}}(x) &= \sup_{r \in \Delta_D} \{T^{\mathcal{I}}(x, r) \otimes d^{\mathcal{I}}(r, r')\} \\
(C \sqcap D)^{\mathcal{I}}(x) &= C^{\mathcal{I}}(x) \otimes D^{\mathcal{I}}(x) & (\forall R.C_1 + \dots + w_k C_k)^{\mathcal{I}}(x) &= w_1 C_1^{\mathcal{I}}(x) + \dots + w_k C_k^{\mathcal{I}}(x) \\
(C \sqcup D)^{\mathcal{I}}(x) &= C^{\mathcal{I}}(x) \oplus D^{\mathcal{I}}(x) & (C \sqsubseteq D)^{\mathcal{I}} &= \inf_{x \in \Delta^{\mathcal{I}}} \{C^{\mathcal{I}}(x) \Rightarrow D^{\mathcal{I}}(x)\} \\
(\neg C)^{\mathcal{I}}(x) &= \ominus C^{\mathcal{I}}(x) \\
(\forall R.C)^{\mathcal{I}}(x) &= \inf_{y \in \Delta^{\mathcal{I}}} \{R^{\mathcal{I}}(x, y) \Rightarrow C^{\mathcal{I}}(y)\}
\end{aligned}$$

A fuzzy interpretation \mathcal{I} *satisfies* (is a *model* of) a fuzzy statement $\langle \alpha, n \rangle$ iff $\alpha^{\mathcal{I}} \geq n$. The notions of logical consequence, best entailment degree and best satisfiability degree of α are as for Section 2. We additionally define the *Best Satisfiability Degree* (BSD) [3] of a concept C w.r.t. a fuzzy KB \mathcal{K} as

$$bsd(\mathcal{K}, C) = \sup_{\mathcal{I} \models \mathcal{K}} \sup_{x \in \Delta^{\mathcal{I}}} C^{\mathcal{I}}(x).$$

MCDM and Fuzzy DLs. We next provide some examples, illustrating how to encode some simple (fuzzy) MCDM problems in fuzzy DLs, showing the potential of our approach.

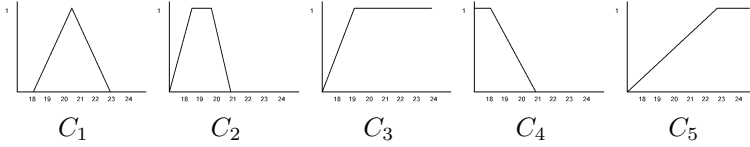
Example 1. The MCDM problem consists of four alternatives and five criteria for an electrical power dispatching system in the case of shortage of electrical power [2]. The four alternatives correspond to four regions of a city to which to give priority. The five criteria correspond to C_1 (Residential area), C_2 (Shopping centers), C_3 (Clubs and recreation centers), C_4 (Educational centers), and C_5 (Medical urgent care centers). We normalize the performance values according to $q_{ij} = w_j \cdot (a_{ij} / \sum_{l=1}^n a_{il})$. The decision matrix together with the matrix of the q_{ij} 's is shown below:

		Criteria				
		0.20	0.1	0.3	0.35	0.05
Alternatives		C_1	C_2	C_3	C_4	C_5
x_1	A_1	0.5	0.7	0.3	0.1	0.3
x_2	A_2	0.2	0.3	1.0	0.7	0.2
x_3	A_3	1.0	0.8	0.5	1.0	0.5
x_4	A_4	0.3	0.2	0.2	0.2	1.0

$$q_{ij} = \begin{pmatrix} 0.0455 & 0.0333 & 0.0391 & 0.0149 & 0.0073 \\ 0.0182 & 0.0143 & 0.1304 & 0.1043 & 0.0049 \\ 0.0909 & 0.0381 & 0.0652 & 0.1489 & 0.0122 \\ 0.0273 & 0.0095 & 0.0261 & 0.0298 & 0.0244 \end{pmatrix}$$

² The use case is inspired by <http://med.ee.nd.edu/MED11/pdf/papers/t3-013.pdf>

Now, we further assume that each of the five criteria has *dynamic, time dependent* electricity demand on the day time (18-24) as depicted below (values are normalized to one).



We model this situation by assuming that we have in the KB the axioms:

$$\begin{aligned}
 C_1 &= \text{ResidentialArea} \sqcap \exists \text{hasDemand.tri}(18, 20, 50, 23) & C_2 &= \text{ShoppingCenter} \sqcap \exists \text{hasDemand.trap}(17, 19, 20, 21) \\
 C_3 &= \text{ClubRecreationCenter} \sqcap \exists \text{hasDemand.rs}(17, 19) & C_4 &= \text{EducationalCenter} \sqcap \exists \text{hasDemand.ls}(19, 21) \\
 C_5 &= \text{MedicalUrgentCareCenter} \sqcap \exists \text{hasDemand.rs}(23, 24)
 \end{aligned}$$

where `hasDemand` is a functional concrete feature. We now define the four alternatives A_i as the following weighted concepts:

$$A_i = q_{i1} \cdot C_1 + q_{i2} \cdot C_2 + q_{i3} \cdot C_3 + q_{i4} \cdot C_4, \text{ for } i = 1, 2, 3, 4.$$

The *final rank value* of alternative A_i w.r.t. a knowledge base \mathcal{K} , denoted $rv(\mathcal{K}, A_i)$ is defined as $rv(\mathcal{K}, A_i) = \text{bsd}(\mathcal{K}, A_i)$, i.e. we compute its maximal satisfiability degree. It is thus, easily verified that this extends the WSM to the fuzzy DL case. Finally, the *optimal alternative* is $A^* = \arg \max_{A_i} rv(\mathcal{K}, A_i)$. Now, assume that we have a shortage of power between time 19-20. It can be verified that (the values have been computed using the fuzzy DL reasoner FUZZYDL [3]) $rv(\mathcal{K}, A_1) = 0.11625$, $rv(\mathcal{K}, A_2) = 0.25628$, $rv(\mathcal{K}, A_3) = 0.28856$, $rv(\mathcal{K}, A_4) = 0.07632$, and, thus, the ranking of the alternatives is $A_3 \succ A_2 \succ A_1 \succ A_4$ and the optimal alternative is $A^* = A_3$. \square

Example 2. The following example is a simplified version of [4] and is about landfill siting. We have to select among two sites, $Site_1, Site_2$, according to two criteria (TI -Transportation Issues, and PN -Public Nuisance) and there are two experts (E_1, E_2). The decision matrix of the experts is shown below:

E_1		Criteria		E_2		Criteria	
		0.48	0.52			0.52	0.48
Alternatives		C_1	C_2	Alternatives		C_1	C_2
x_1	A_1	$tri(0.6, 0.7, 0.8)$	$tri(0.9, 0.95, 1.0)$	x_1	A_1	$tri(0.55, 0.6, 0.7)$	$tri(0.4, 0.45, 0.5)$
x_2	A_2	$tri(0.6, 0.7, 0.8)$	$tri(0.4, 0.5, 0.6)$	x_2	A_2	$tri(0.35, 0.4, 0.45)$	$tri(0.5, 0.55, 0.6)$

Note that this time, the performance of the alternatives is defined in terms fuzzy numbers. We may model the scenario as follows. For each expert $k = 1, 2$, for each alternative $i = 1, 2$ and for each criteria $j = 1, 2$, we define the concept

$$P_{ij}^k = \exists \text{hasScore}.a_{ij}^k,$$

where `hasScore` is a concrete feature and a_{ij}^k is, according to expert k , the performance of alternative i with respect to criteria j (a_{ij}^k is the fuzzy number in the matrix). Now, for each expert k and alternative i , we define the weighted concept

$$A_i^k = w_1^k \cdot P_{i1}^k + w_2^k \cdot P_{i2}^k,$$

which takes into account also the weight w_j^k of expert k for criteria j . Finally, we combine the two experts outcome, by defining the weighted concept

$$A_i = 0.5 \cdot A_i^1 + 0.5 \cdot A_i^2 .$$

Note that we rate both experts equally (this may be changed, of course). The final rank value of alternative A_i w.r.t. a knowledge base \mathcal{K} and the optimal alternative is determined as for Example 1. It can be verified (using again the fuzzy DL reasoner FUZZYDL 3) that $rv(\mathcal{K}, A_1) = 0.26$ and $rv(\mathcal{K}, A_2) = 0.37$ and, thus, the ranking of the alternatives is $A_2 \succ A_1$ and the optimal alternative is $A^* = A_2$. \square

5 Conclusions

We have made an initial step in addressing MCDM within fuzzy DLs and, thus, towards a fuzzy knowledge-assisted approach to decision making. Our aim here was exploratory on the argument and a more in depth investigations need to be addressed, of course.

There are several points that may be of interest for future research: (i) each alternative is indeed a fuzzy set and, so far, we order alternatives according to the best satisfiability degree. Other methods can be explored by relying on the fuzzy membership function of the alternatives, e.g. using defuzzification methods; (ii) fully exploit fuzzy numbers as performance and weight values in decision matrixes; (iii) for illustrative purposes, we have just considered a simple, though widely used, basic MCDM method, namely the weighted sum method. The MCDM literature (inclusive their fuzzy MCDM variants) is quite large, so it will be of interest whether and how other methods can be integrated within fuzzy DLs as well. E.g., for illustrative purposes, we considered the weighted sum method, though in general other fuzzy aggregation operators may be needed to combine the multiple performance values into an aggregated value (e.g., to cope with criteria interdependence/conflict); (iv) exploit the fact that we may express background/domain knowledge within fuzzy DLs (e.g., in Example 1 we may include an Urban Ontology to formalize the problem, which is part of larger GIS system; a similar argument applies to Example 2 as well); (v) fuzzy DLs are parametric with respect to t-norm, t-conorm, etc. Choosing, e.g., appropriate fuzzy connectors, as well as appropriate fuzzy aggregation operators, is clearly application specific and may bring to different results and, thus, these issue needs both theoretical and empirical investigations in our setting as well.

References

1. Baader, F., Calvanese, D., McGuinness, D., Nardi, D., Patel-Schneider, P.F. (eds.): The Description Logic Handbook: Theory, Implementation, and Applications. Cambridge University Press, Cambridge (2003)
2. Bobillo, F., Delgado, M., Gómez-Romero, J.: Delorean: A reasoner for fuzzy OWL 1.1. In: Proceedings of the 4th International Workshop on Uncertainty Reasoning for the Semantic Web (URSW 2008). CEUR Workshop Proceedings, vol. 423, p. 10 (2008)

3. Bobillo, F., Straccia, U.: fuzzyDL: An expressive fuzzy description logic reasoner. In: International Conference on Fuzzy Systems (FUZZ 2008), pp. 923–930. IEEE Computer Society, Los Alamitos (2008)
4. Chang, N.B., Parvathinathan, G., Breeden, J.B.: Combining GIS with fuzzy multicriteria decision-making for landfill siting in a fast-growing urban region. *Journal of Environmental Management* 87(1), 139–153 (2008)
5. Hájek, P.: *Metamathematics of Fuzzy Logic*. Kluwer, Dordrecht (1998)
6. Kahraman, C.: *Fuzzy Multi-Criteria Decision Making: Theory and Applications with Recent Developments*. Springer, Dordrecht (2008)
7. Klir, G.J., Yuan, B.: *Fuzzy sets and fuzzy logic: theory and applications*. Prentice-Hall, Inc., Upper Saddle River (1995)
8. Lukasiewicz, T., Straccia, U.: Managing uncertainty and vagueness in description logics for the semantic web. *Journal of Web Semantics* 6, 291–308 (2008)
9. Stoilos, G., Simou, N., Stamou, G., Kollias, S.: Uncertainty and the semantic web. *IEEE Intelligent Systems* 21(5), 84–87 (2006)
10. Straccia, U.: Reasoning within fuzzy description logics. *Journal of Artificial Intelligence Research* 14, 137–166 (2001)
11. Straccia, U.: Description logics with fuzzy concrete domains. In: Bachus, F., Jaakkola, T. (eds.) 21st Conference on Uncertainty in Artificial Intelligence (UAI 2005), Edinburgh, Scotland, pp. 559–567. AUAI Press (2005)
12. Straccia, U.: A fuzzy description logic for the semantic web. In: Sanchez, E. (ed.) *Fuzzy Logic and the Semantic Web. Capturing Intelligence*, ch. 4, pp. 73–90. Elsevier, Amsterdam (2006)
13. Triantaphyllou, E.: *Multi-Criteria Decision Making Methods: A Comparative Study*. Kluwer Academic Publishers, Dordrecht (2000)

A Hybrid System Combining Description Logics and Rules for Inventive Design

Alexis Bultey, Cecilia Zanni-Merk, François Rousselot, and François de Beuvron

LGECO – INSA de Strasbourg
24 bd de la Victoire – 67084 Strasbourg – France
{alexis.bulteuy,cecilia.zanni,francois.rousselot,
francois.debeuvron}@insa-strasbourg.fr

Abstract. Knowledge acquisition and capitalization to solve problems concerning artefact evolution, still called inventive design, has a certain quantity of specific characteristics. The theoretical approach we are interested in, called TRIZ (the Russian acronym for Theory for Inventive Problem Solving), when translated into a methodological procedure, can be declined into two different steps: problem formulation and problem resolution. This article presents an analysis of two of the most used knowledge bases of TRIZ during the resolution stage. These knowledge bases have been formalized by the construction of an ontology of the informal knowledge sources usually used by the TRIZ experts. This approach has permitted the design of a software architecture that eases the implementation of these bases by means of their declarative manipulation. It combines rules and description logics for populating the ontology and facilitates the access to the compiled generic knowledge that synthesizes, at an abstract level, the already encountered problems and their solutions.

Keywords: Knowledge engineering, Ontologies, Inventive design, Knowledge-based systems, Knowledge capitalization.

1 Introduction

The inventive design theory TRIZ [1, 2] shares some close concepts with the Artificial Intelligence field. It is to be remarked that TRIZ can be declined in a set of tools and methods that are based on concepts that are not formally defined.

Since several years now, we have been working in this way and our aim is to clarify the fuzzy points to have a better comprehension of the methodology, of the manipulated knowledge and the specific TRIZ reasoning. Our works have concerned the modelling of the formulation process and the problem solving process [3].

In this article, we will focus on the exploitation of two of the TRIZ knowledge bases that are used during the resolution phase. We will also describe a software architecture that we have designed for implementing them. Our goal is to ease the exploitation of those knowledge bases and to permit their extension by the capitalization of new knowledge, for example.

In fact, Altshuller, the TRIZ creator, synthesized the knowledge that appeared recursively in patents. His proposal is simple, he has remarked that the ideas behind a patent could often be useful for solving a problem belonging to other field.

In this way, a simple procedure, such as heating under pressure followed by an abrupt cooling and depression may be used for solving different problems: coring sweet peppers, opening cedar nuts, detaching sunflower seeds, producing castor sugar, cleaning filters, splintering imperfect crystals.

Initially, the capitalization of the knowledge extracted from almost 40 000 patents, produced the Contradiction Matrix, where 40 inventive principles were statistically extracted from patents [1, 2]. But this approach was not enough. That is the reason why Altshuller, from 1973 on, has developed a more useful conceptual tool, the Substance-Field Analysis (SFA), which is used with the pointers to scientific-engineering effects.

In this article, we are interested in the formalization of the Substance-Field Analysis (SFA) and the pointers to scientific-engineering effects. In first section, we briefly present their related notions, and we describe their modeling in the second section. In the end, we discuss the way our approach has permitted the design of a software architecture that manipulates this knowledge in a declarative way and facilitates the integration of new knowledge.

2 The TRIZ Main Notions

In our previous works, we have defined an ontology that covers the majority of the notions of TRIZ and, mainly, the elements that permit the formalization of the different models as they are used in inventive design [3].

Next subsections summarize the notion of the SFA and pointers.

2.1 Substance-Field Analysis (SFA)

To model the physical structure of a problematic system, the so-called SFA is used. The basic idea behind this analysis is that any problematic part of an engineering system can be represented as a set of substance components and field interactions between those components.

To obtain a solution to the problem means that the given physical structure which contains the undesirable interaction has to be transformed into a structure in which the desired interaction is achieved. Inventive Standards are the rules which indicate what patterns are to be used to transform a given substance-field model of a problem into a solution model.

Standards are built in the form of recommendations, and generally, formulated as rules like *If < Condition1 > and < Condition2 > then < Recommendation >*. Both conditions permit recognizing the typology of the problem associated to the standard. This way, for a given problem model, there exist a certain number of recommendations allowing the construction of the corresponding solution model.

In TRIZ, 76 Inventive Standards are available. A problem with these Inventive Standards is that they are formulated abstractly, so their practical use is quite difficult.

Example – Inventive Standard 1.2.2: If there are useful and harmful effects between two substances and it is not required that these substances be closely adjacent to one another, but it is forbidden or inconvenient to use a foreign substance, the problem is solved by introducing a modification of the existing substances.

2.2 Pointers to Scientific-Engineering Effects

Another TRIZ tool is the collection of so-called scientific-engineering effects. While Inventive Standards do not produce recommendations in terms of what physical substances or fields should be used, scientific-engineering effects provide the mapping between TRIZ technical functions and known natural phenomena.

Example: Displacement of an object for a short distance, apply the effect of thermal expansion to control it.

TRIZ also contains unique collections of geometrical and chemical effects which relate chemical and geometrical knowledge with technical functions.

3 The Representation Choice: Ontologies

With the idea of designing an architecture for problem solving, we have represented the SFA by means of an ontology and we have decided to study a new way of representing the pointers to scientific-engineering effects. The goal is to make their utilisation more flexible and to promote their extension.

The methodology we have followed for building the ontology is based on the properties classification criteria proposed by [4]. Our modelling works are based on a reference text corpus [1, 2, 5, 6] and on discussions with several TRIZ experts. These works have given birth to an ontology (full version can be found in [11]). The logical coherence is assured by a Description Logics (DL) based system, CICLOP [8].

As DL need an upper layer for the organization of the different steps for the building of the resolution model, we have developed an environment that embodies it. With the goal of preserving the declarative characteristics of the whole implementation, we have chosen a rule based system, JSNARK, which communicates with CICLOP. JSNARK is a JAVA implementation of a rule based system, close to SNARK [13]. This is a development by François Rousselot, one of the co-authors of this article. This engine permits the inspection of the concepts in the knowledge base, the asking of questions to the user and the launching of classifications anytime. Our knowledge base is decomposed into two terminological knowledge bases (TBox) contained in CICLOP and an inferential knowledge base in JSNARK.

The TBoxes describe, on the one hand, the knowledge about the SFA; and on the other hand, their transformations, as suggested in the Inventive Standards and the pointers to the scientific-engineering effects. They constitute the Su-Field ontology and the Transformations ontology. They are, evidently linked, but we are going to present them separately for clarity reasons.

The inferential base concerns knowledge about the reasoning mechanisms described in the Inventive Standards (the IS rules), the application rules (establishing when a certain Inventive Standard may be applied, for example), and the project rules that allow project management.

3.1 The Su-Field Terminological Base

It defines the “active” concepts in the SFA (cf. figure 1). Some of them are essentials for the problem modeling stage (Field, Substance). Some other play an important

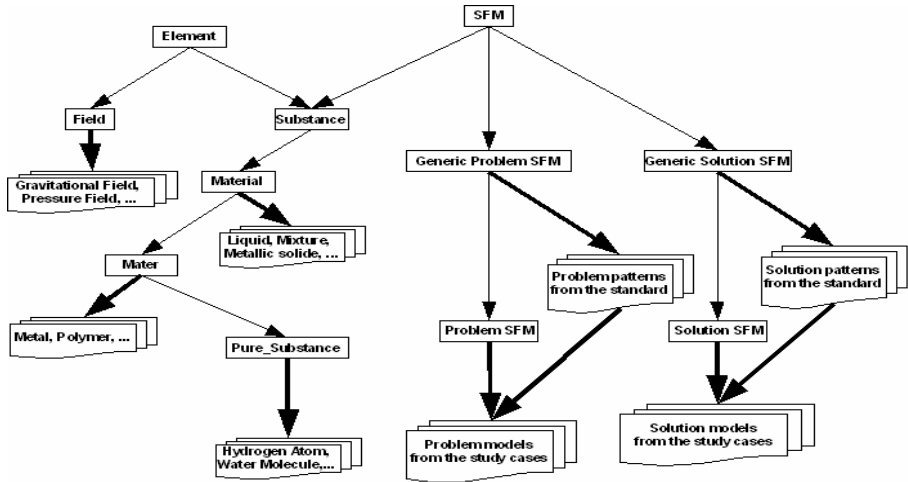


Fig. 1. Selection of the Su-Field ontology

role in the standard application stage (Problem & Solution patterns). SFM is the top concept of the Su-Field TBox.

A SFM is defined as a TRIZ model that concerns at least an Element. His direct sub-concepts are the generic problem SFM and the generic solution SFM. They represent the different patterns of problem and solution in form of SFM and they are extracted from the standards. Under those concepts, we introduced the sub-concepts of problem SFM and solution SFM. A problem SFM categorises the different problems that have been modelled by a designer in the form of a SFM. In the same way, the solution SFM corresponds to the model of solution obtained thanks to the application of an appropriated standard to a particular problem.

Element describes the two components (substance and field) of a substance-field model. The son concepts of Element are, therefore, Substance and Field and they are defined as being disjoint. The son concepts of Field characterize the so-called “technological” fields that we find in TRIZ. As in TRIZ a substance may be defined as a substance-field model, its concept is, therefore, subsumed by the SFM and Element concepts. For example, a magnet may be modelled as a SFM considering a ferromagnetic substance and a magnetic field.

There are two levels of substance decomposition: the macro-level is represented by the concept of Material and the micro-level is represented by the concept Matter. Matter refers to a specific substance composed of some substances categorized as Pure Substance (atoms, molecules) and composed of some fields named Bonding Field (hydrogen bonding field, metallic bonding field ...). Material is composed of Matter, and a “technological field”; it has an attribute specifying its physical state (solid, liquid, gas, plasma).

3.2 The Transformations Terminological Base

It defines every transformation as a pair of SFM; the first corresponds to the description of the initial state and the other, the final state. This ontology classifies all the transformations stated in the Inventive Standards and their applications to 40 particular problems. This ontology contains also 200 physical effects that were considered as innovative by Altshuller (cf. figure 2).

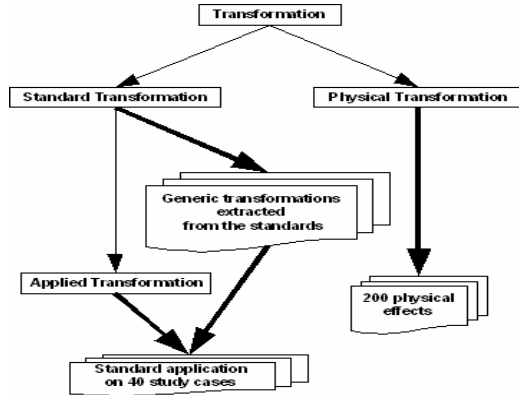


Fig. 2. Selection of the Transformations ontology

The standard transformation represents the generic transformation performed by the standards. It considers a generic problem SFM as initial state and a generic solution SFM as final state. The applied transformation (a direct son concept of standard transformation) corresponds to the application of standards to a specific problem. It considers a pair of specific SFMs: a problem SFM and a solution SFM. Finally, the physical transformation regroups the physical effects in the form of a SFM transformation. Here, the sub-concepts of substance and field are intensively used to precisely express the transformation produced by an effect.

The main role of this knowledge base consists in capitalizing the last SFA and the physical effects which were used to resolve particular problems.

3.3 The Inferential Base

Contained in JSNARK, it is organized in 2 hierarchical levels: The project rules and the IS (standing for Inventive Standards) rules.

The project rules supervise and articulate the different stages of the SFA: the modelling problem stage, the resolution stage and the searching of similar standard or physical transformations stage. The first stage rules propose to the user to specify the elements of his problematic SFM (substance, field, interaction,...) by given him a list of specific elements (metallic solid, magnetic field, useful interaction...) available in the Su-Field TBox. They enable the construction of a

new axiom of problem SFM which is classified in the SFM TBox and associated to a generic problem SFM as a direct subconcept thanks to the subsumption inference. The resolution stage rules control the application of the **IS rules** and enable the user to add some constraints in the solution SFM if it's necessary.

The IS rules are the formalization of the 76 standards. Those rules propose a systematic way to apply standards, contrary to the confusing traditional approach. They considered in premise a generic problem SFM and in conclusion a generic solution SFM. Moreover, they assist the designer in specifying the generic solution SFM in a solution SFM by asking questions about the problem context.

Finally, the searching stage rules launch some matching algorithms to find some similar standard applications and similar physical transformations. Those algorithms are based on semantic matching [12]. They intend to propose to the user :

- some old problems which have been transformed by the same standard,
- some effects which could physically enable the needed transformation.

4 An Example

For understanding the general operation of this architecture for inventive problem solving, we present it on a concrete problem (cf. figure 3). This example is one of 40 problems used to validate our system. They were extracted from the TRIZ literature, and we have compared the solutions obtained from the traditional approach to the solutions given by our system in order to evaluate it. It is a pumping problem: "The pumping of a liquid through a pipe is often a source of problems. The fact that a gate or a valve interrupts the liquid flow often produces pressure changes in fits and starts that may destroy the pipe".

The SFM associated with the problem contains two substances: the pipe and the liquid. It contains also two fields: a mechanical field providing a useful interaction (to contain the water) and a field of impact causing a harmful interaction (to destroy the pipe) between the two substances.

The Pumping Problem SFM is incorporated into the Su-Field TBox, translated into description logics thanks to the project rules (1). This SFM is then classified by the DL engine in order to be associated to a generic model SFM (useful/harmful simple SFM) (2). Then, the IS rules are launched, they ask some additional questions to the user in order to converge to a generic solution SFM (Internal developed simple SFM) (3). The project rules specify this generic solution SFM according to the problem SFM, and they introduce a new solution SFM (4). The project rules combine these new problem and solution SFMs to create a new applied transformation (Applied transformation to Pumping) (5). Once this applied transformation is classified in the Transformation TBox, the searching for similar applied transformations is performed (6) as well as the searching of relevant physical effects (7). If these searchings are successful, the project rules propose the results to the user (8 & 9) by mean of a graphical interface [13].

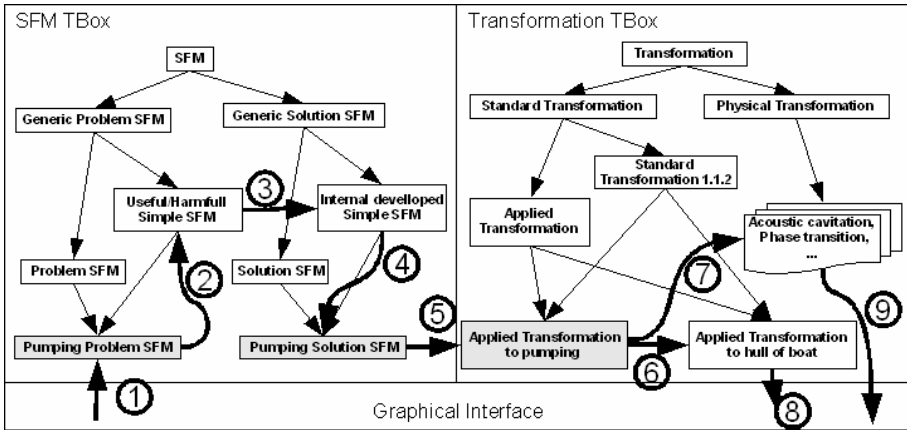


Fig. 3. Selection of the Transformations ontology

5 Conclusions

The Su-Field ontology and the Transformations ontology are a considerable contribution to the necessary normalization of the TRIZ knowledge. They formally connect the Scientific-Engineering effects and the Inventive Standards that were, until now, used in a separate way during the Substance-Field Analysis.

We have chosen to use description logics for managing these ontologies. The interest of this declarative formalism, beyond the coherence control, is that we can, on the one hand, make evident the continuum of abstraction levels among the more or less acquainted models; and, on the other hand, take into account any modification or addition of concepts by the propagation of the introduced constraints.

Moreover, in addition to SFA, different questioning strategies are now conceivable, whereas the strategy was fixed according to the Inventive Standards. We have the possibility, for example, of asking at the beginning of the process, various security questions (such as, is there any forbidden substances?) The knowledge representation system, by the propagation of the known true facts, will automatically exclude the forbidden substance in the searching of effects.

This implementation in a description logics system permits different ways of consulting the effects knowledge base. It is possible to propose an effect and study the transformations where it can be involved.

Regarding the extensibility of this method, we observe that the higher level ontologies do not presuppose that the nature of the SFM be material. Therefore, we may take in account the possibility of solutions in non-technical fields, for example, in organizational, communicational or software engineering problems.

As for the possibility of capitalizing new knowledge (for example, the addition of a new physical effect), it is enough to describe it with the ontology terms and CICLOP will be able to automatically classify it. The possibility of capitalization of new knowledge of TRIZ type is now open; there is an enormous work in perspective, consisting in the addition of knowledge about non-technical fields.

References

- [1] Altshuller, G.S.: Creativity as an exact science. Gordon and Breach, New York (1988); 0275-5807
- [2] Altshuller, G.S.: TRIZ the innovation algorithm, systematic innovation and technical creativity. Technical innovation Center Inc., Worcester, Massachusset (1999)
- [3] Zanni, C., Cavallucci, D., Rousselot, F.: An Ontological Basis for Inventive Design, Computers in Industry. Special issue on "Computer Aided Design" (to be appeared, 2009)
- [4] Guarino, N., Welty, C.: A formal ontology of properties. In: Dieng, R., Corby, O. (eds.) EKAW 2000. LNCS (LNAI), vol. 1937, pp. 97–112. Springer, Heidelberg (2000)
- [5] Savransky, S.D.: Engineering of Creativity: Introduction to TRIZ methodology of Inventive Problem Solving. CRC Ed., Boca Raton (2000)
- [6] Salamatov, Y.: TRIZ: the right solution at the right time (2000); Insystec Ed. Translated by Strogaia M., Yakovlev S.
- [7] Bultey, A., De Bertrand De Beuvron, F., Rousselot, F.: A problem solving environment based on TRIZ ontologies. In: Proceedings of Virtual Concept 2006, Playa del Carmen, Mexico (2006)
- [8] De Bertrand De Beuvron, F., Rousselot, F., Kullman, M., Rudolf, D.: CICLOP. In: Proceedings of « les journées francophones de l'ingénierie des connaissances », Toulouse, France (2000)
- [9] Lauriere, J.L., Vialatte, M.: SNARK: a language to represent declarative knowledge and inference engine which use heuristics. In: Proceedings of IFIP Congres, Dublin, Ireland, pp. 811–816 (1986)
- [10] Nardi, D., Brachman, R.J.: An Introduction to Description Logics. In: Baader, F., Calvanese, D., McGuinness, D.L., Nardi, D., Patel-Schneider, P.F. (eds.) The Description Logic Handbook, pp. 5–44. Cambridge University Press, Cambridge (2002)
- [11] Bultey, A., De Bertrand De Beuvron, F., Rousselot, F.: A substance-field ontology to support the TRIZ thinking approach. *Int. J. Computer Applications in Technology* 30(1/2), 113–124 (2007)
- [12] Baader, F., Küsters, R.: Matching in description logics with existential restrictions. In: Proc. of the 7th Int. Conf. on Principles of Knowledge Representation and Reasoning (KR 2000), pp. 261–272 (2000)
- [13] SAAS, <http://www2.insa-strasbourg.fr/lgeco/these/>

Domain Modeling Based on Engineering Standards

Carlos Toro¹, Manuel Graña³, Jorge Posada¹, Javier Vaquero¹, Cesar Sanín²,
and Edward Szczerbicki²

¹ VICOMTech Research Centre, Spain

² Faculty of Engineering and Built Environment, University of Newcastle, Australia

³ University of the Basque Country - Facultad de Informática de San Sebastián, Spain

Abstract. In this paper we present a new methodology for Domain modeling based on Engineering Standards. We discuss some benefits of standards as guidelines for a Knowledge Based Domain modeling, potential challenges and approaches to overcome them. The benefits of using Standards as models for Domain ontologies have been shown as valid in related work and, as proof of concept, we present a case study where our methodology was successfully applied.

Keywords: Domain Modeling, Knowledge Based Systems, ontologies.

1 Introduction

Webster defines a Domain as “a sphere of knowledge, influence or activity”. In Computer Science, we consider a Domain as a sphere of Knowledge identified by a name, in which the information is a collection of concepts, intermediations and facts about entities [8]. In other words, a Domain describes the elements and characteristics belonging to a Knowledge Base. A Domain ontology (or Domain-specific ontology) models the Knowledge in a specific Domain, representing particular meanings for the terms it contains. For example, in the Plant Design Domain, the concept *elbow* is a specific type of bend pipe used to change the direction of the fluid. However, when considering similar Domains, the degree of specialization and conceptualization, even within the same concept, can vary slightly (e.g. a Piping Engineer will consider a wider definition of an elbow when compared to a Structural Engineer). The Domain modeling does not stop at the concept definition. Any concept in a Domain also needs property characterization, e.g. for an elbow in the Plant Design Domain, distinctiveness such as the radius, the curvature, etc. According to BSI [1], a standard is an agreed, repeatable way of doing something. It is a published document that contains a technical specification or other precise criteria designed to be used consistently as a rule, guideline, or definition. A comprehensive list of standards organizations can be found in [5]. In this paper, we will present a methodology that can be used to aid in the Domain modeling of a Knowledge Base using Engineering Standards. According to BSI [1], a standard is an agreed, repeatable way of doing something. It is a published document that contains a technical specification or other precise criteria

designed to be used consistently as a rule, guideline, or definition. Standards help to make life simpler and to increase the reliability and the effectiveness of many goods and services we use. Standards are created by bringing together the experience and expertise of all interested parties such as the producers, sellers, buyers, users and regulators of a particular material, product, process or service. This paper is structured as follows: In section 2, we will present a state of the art on the different topics relevant to our work. In section 3, we present our methodology for Domain modeling based on Engineering Standards. In section 4, we present a case study where we successfully applied our methodology, and lastly in section 5, we present some conclusions and future work.

2 State of the Art

In this section we introduce some concepts relevant to this paper. Knowledge is considered an invaluable resource of great benefit for most purposes in life. For this reason, mankind has always attempted to make it part of their assets. Knowledge itself seems to be an attribute of human beings; it may be defined [7] as: *(i)* the expertise and skills acquired by a person through experience or education via a theoretical or practical understanding of a subject, *(ii)* what is known in a particular field related to facts and information or *(iii)* experiential knowledge, the awareness or familiarity gained by experience of a fact or situation. Knowledge Engineering (KE) is an engineering discipline that involves integrating Knowledge into computer systems in order to solve complex problems, normally requiring a high level of human expertise [2]. Following this line of thought, Knowledge Bases can be modeled and used by computer systems in order to enhance their capacities. One of the most commonly used techniques for Knowledge modeling is ontologies.

2.1 Knowledge Modeling Using Ontologies

We base our approach on the widely accepted definition of ontology given by Tom Gruber of what an ontology is in the Computer Science Domain: an ontology is the explicit specification of a conceptualization; in other words it is a description of the concepts and relationships in a Domain [3]. Some of the reasons to use ontologies in Knowledge modeling are: *(i)* To separate Domain Knowledge from actual Knowledge, *(ii)* to analyze Domain Knowledge, *(iii)* to share common understanding of the structure of information between people or software agents, *(iv)* to enable reuse of Domain Knowledge, and *(v)* to make Domain assumptions explicit. To our knowledge, there are few reported cases where Standards are used along with semantic technologies, such as the notorious case of CIDOC-CRM [7], whose primary role is as a formal ontology intended to facilitate the integration, mediation and interchange of heterogeneous cultural heritage information heterogeneous sources. The usual approach for modeling Domain starts with human experts who use their own knowledge about the specific needs of an industry in order to model the subjects and their

relations using editors or even plain paper. In some companies the Domain expert is known as the “Knowledge Engineer” and their purpose is to conceptualize the business now-how and processes.

3 A Methodology for the Use of Engineering Standards as Models for Domain

In this section, we introduce our methodology for the Domain modeling based on Engineering Standards.

3.1 Motivation to Use Standards as a Basis for Domain Modeling

The ability to demonstrate compliance with industry standards is an effective mean of differentiation in a competitive marketplace. In addition, manufacturing products or supplying services to appropriate standards maximizes their compatibility with those manufactured or offered by others, thereby increasing potential sales and widespread acceptance. As consumers become increasingly informed about their choices, conformity to recognized standards becomes pivotal. We argue that the use of Engineering Standards as models for Domains provides the following benefits:

- **Consensus:** There is a consensus on the terminology, organization and logic of the Domain.
- **Format support:** Many Engineering applications support Engineering Standards as input/output formats (e.g. CAD software with STEP modules [7]). This fact helps in the categorization of elements and the mapping of such elements into the Knowledge Base).
- **Avoidance of Semantic loss:** The modeling of an Engineering Standard, usually considers not only of the element in its isolated form, but also, the relationship of the element with surrounding objects. The aforementioned fact is indeed a valuable feature that helps in the conservation of the Semantics properties of such elements.
- **Ease of a new Domain modeling based on existing Standards:** If there is no existing Engineering Standard for a given Domain, a standard complying with similar characteristics can be used. An example of this is the use of STEP application protocol 227 (Plant Design) ([7], [8]) in the case of Ship Design.
- **Standards are revised on a regular basis:** Their nature is intrinsically evolutionary due to the development of new technologies for fabrication or the typical evolution of engineering paradigms. When using standards as a basis for a Knowledge Base, there exists an intrinsic guarantee that the most recent data models will be used.

Our methodology is divided into a series of logical steps that must be performed to assure a correct modeling of the Domain (see Fig. 1).

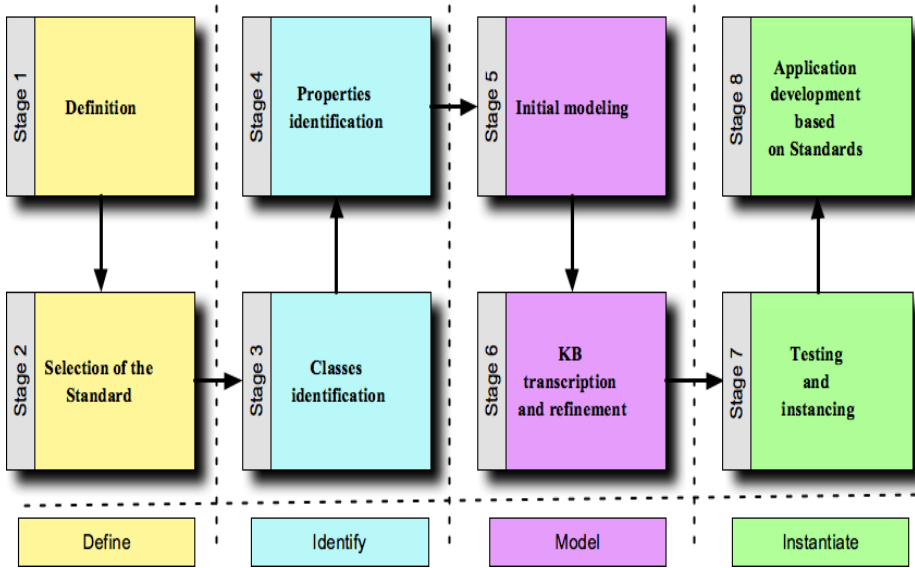


Fig. 1. Our methodology for Domain modeling based on Standards

As can be seen, we divided our methodology in 4 layers: DEFINE, IDENTIFY, MODEL and INSTANTIATE, encompassing eight stages as follows:

<p>DEFINE:</p> <ul style="list-style-type: none"> • Stage 1- Definition: In this stage, an identification of the purpose and requirements of the domain is made. What is the purpose of the KB, the information that will be stored and the needed level of detail of such information. • Stage 2- Selection of the Standard In this stage, there is a search of a standard that suits the defined needs. As a result of this selection, the chosen standard must be studied in detail, how is constructed, what can be done in order to extend it, etc
<p>IDENTIFY:</p> <ul style="list-style-type: none"> • Stage 3- Class identification In this stage, an identification of the purpose and requirements of the domain is made. What is the purpose of the KB, the information that will be stored and the needed level of detail of such information. • Stage 4- Property identification At this stage, the characteristics that can be measured or determined by data types (string, Boolean value, integer, etc), are identified in each class, e.g. outside diameter, length, etc. Then we identify the characteristics that relate a class with other classes (relation types). In general data type characteristics are easily recognizable and obtained by simple interrogations (e.g. to a geometric model). Relation types are a little bit more difficult to find, because generally when talking about a geometric model whose characteristics will be obtained, those sets of elements are categorized as geometric primitives rather than functional objects. For the aforementioned case, solutions like the process of branding and matching presented by Posada, could be used [4].
<p>MODEL:</p> <ul style="list-style-type: none"> • Stage 5- Initial modeling In this stage, a subset of the domain is chosen in order to verify the complexity of the overall modeling and the real capabilities of the KB. Since the elaboration of a KB is an iterative process by nature, this small test must answer initial modeling needs. • Stage 6- KB transcription and refinement Sometimes the initial modeling is enough for the KB to fulfill the design requirements in stage 1, however is highly recommended to perform a verification of the transcription, by using for example the capabilities of an ontology reasoner to check the congruence of the KB. Once the transcription is done, a refinement process takes place. In such stage any needed extension of the standard take place. Usually the transcription a refinement is performed using an editor.

INSTANTIATE:**• Stage 7- Testing and instancing**

In this step the test of the instances and the creation of an automatic or semiautomatic instancing mechanism (if needed) is performed. As a final step, some individuals conforming to the specification of the classes can be manually modeled using again an editor. Such process can be automated if needed if any API tools are available and the elements that must conform can be interrogated e.g. modeling an industrial plant KB that has a 3D counterpart in CAD, the CAD API can be used to interrogate the elements and the editor API to semi-automatically “fill” the individuals in the KB. This step does not strictly fall into the modeling process, but in order to really use the modeled KB is needed.

• Stage 8- Application development based on Standards

In this last step the Virtual Engineering Application using the Domain model is developed, this last stage comprises the actual usability of the Domain where the VEA advantages from Semantics via the enhancement obtained by having a better described and consensual Domain model.

3.2 Potential Challenges

There are some potential challenges to take into account when modeling a Domain based on Engineering Standards, for instance: *(i)* The design of the Standard could be Functionally biased: In some cases, the Engineering Standards is functionally oriented, a fact that leads to potential semantic loss, as the Standard does not include the required parameters for a complete Domain modeling. This case is exemplified by the STEP element “Valve” [7]. With this element, the standard reveals only functional parameters (actuator_type, operation_mode, type), but ignores geometric parameters that are needed and also easy read from a CAD model (diameter, length, etc.). In this case an extension of the class should be performed in order to obtain a complete Knowledge Base. When the extension of the Domain is needed, it is advisable to double check if the parameter is a fundamental characteristic. At times, the Engineering Standards offer a way to obtain the required parameter by interrogating neighboring elements (in the case of the Valve, the input and output pipes could be used for such a purpose). If the parameter needed is fundamental and the extension is unavoidable, it should be clearly specified as being “outside the standard feature” and it must follow the ES architecture, e.g. it should be part of the correct element and moreover derived from the correct parent class. *(ii)* The Standard can disappear or be absorbed by other standards: Perhaps due to a lack of use or for administrative reasons, some standards disappear (e.g. the case of CAM-I AIS); in such cases the use of a Knowledge Base based on such Engineering Standards could be continued, however it would be advisable to migrate the Knowledge Base to a new paradigm when available. In the case of absorption by other standards, it is advisable to review the model in order to check its robustness. *(iii)* The Standard falls short for the Domain needs: This indicates a possible immature Engineering Standards, or an inappropriate selection by the Domain Designer. In both cases, it is advisable that a complete reading and understanding of the Standard and an extensive review of the problem’s characterization (requisites).

4 Case Study

In this section we will describe an example of our proposed methodology for Domain modeling based on Engineering Standards.

Stage 1- Definition. Let us consider the problem of modeling an Industrial Plant, and as a practical example, a Flange element.

Stage 2 - Selection of the Standard: By performing an Internet search, we find that there is an ISO Standard that could be used for our needs; this case is ISO 10303 AP 227 (Industrial Plants) [8].

Stage 3 - Classes identification: Consulting the standard we find that a description of a Flange element exists; this description is depicted in Fig. 2.

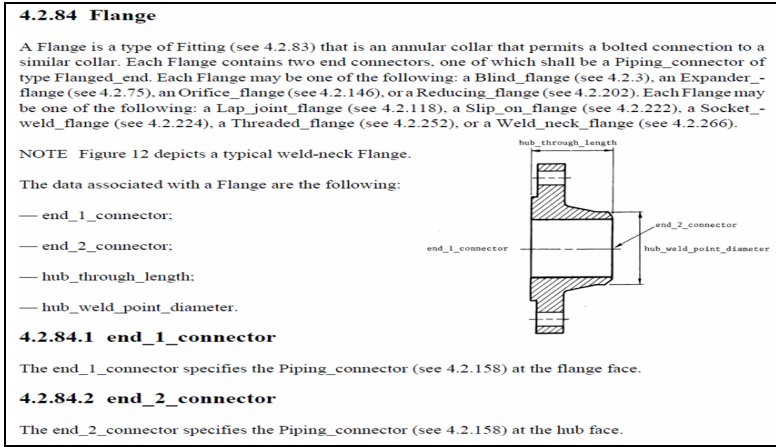


Fig. 2. STEP Excerpt for the case study

Stage 4 - Properties identification. By looking at the properties in the Flange element, we create a classification as can be seen in Table 1. The criterion to choose the concept is provided by an expert in the field whose work as a Knowledge Engineer is fundamental at this stage.

Table 1. Properties classification

Name	Property_type	Value
hub_through_length	Data	Double
hub_weld_point_diameter	Data	Double
end_1_connector	Relational	Element
end_2_connector	Relational	Element

Stage 5 - Initial modeling. We use the Protégé ontology editor to model the element as can be seen in Fig. 3,4, and 5.

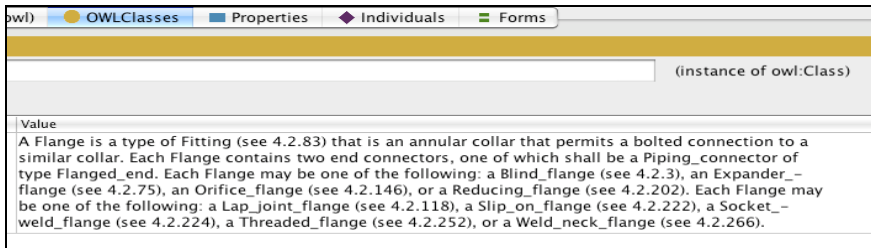


Fig. 3. Modeling of the Flange class

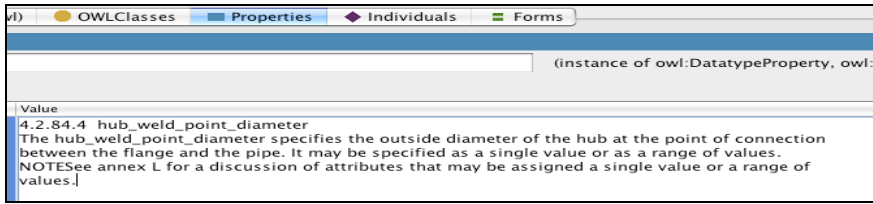


Fig. 4. Modeling of the Flange, data type properties

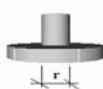



Fig. 5. Modeling of the Flange, relational properties

Stage 6- Knowledge Base transcription and refinement. For the example case, we decide that the Standard contains enough information for our modeling needs, hence no extension is needed. The process is finalized by running a reasoning process to check the ontology for any problems at a logical level (not shown here).

Stage 7- Testing and instancing. In this case, we use the Protégé OWL API for the generation of Java source code suitable for the semi-automatic instancing of individuals, Fig. 6 (left) depicts a section of this code.

Stage 8- Application development based on Standards. As pointed out before, this last stage comprises the actual usability of the Domain. In our case we used the modeled domain in order to match graphic elements coming from a 3D model with their parameters for a semantic synonym graphical adaptation as explained in [8]. Fig. 6 (Right), depicts such matching for the example.

3D CAD (Geometric LOD)	Semantic Synonym representation (parametric)	ISO-STEP Matched parameters
ISO-STEP 10303-AP227 FLANGE (COD. 4.2.84)		
	 <p> r = radius s = side XYZ = Coord. System P = position (px,py,pz) </p>	<p>STEP 4.2.84.3 Hub through length = s</p> <p>STEP 4.2.84.4 hub weld point diameter = $2*r$</p> <p>STEP piping connectors : give XYZ, P.</p>

```

public interface Flange extends OWLIndividual {
    // Property http://www.owl-ontologies.com/Ontology1238934611.owl#end_1_connect
    Pipe getEnd_1_connector();
    RDFProperty getEnd_1_connectorProperty();
    boolean hasEnd_1_connector();
    void setEnd_1_connector(Pipe newEnd_1_connector);

    // Property http://www.owl-ontologies.com/Ontology1238934611.owl#end_2_connect
    Collection getEnd_2_connector();
    RDFProperty getEnd_2_connectorProperty();
    boolean hasEnd_2_connector();
    Iterator listEnd_2_connector();
    void addEnd_2_connector(Pipe newEnd_2_connector);
    void removeEnd_2_connector(Pipe oldEnd_2_connector);
    void setEnd_2_connector(Collection newEnd_2_connector);
}

```

Fig. 6. (Left) Generated control code, (Right) Application development

5 Conclusions and Future Work

In this paper we presented a new methodology for Domain modeling based on Engineering Standards. We discussed some of the benefits of standards as guidelines for a Knowledge Based Domain modeling and some potential challenges along with possible approaches to overcome them. As future work, we intend to test and compare our methodology and other established methods that could be used for domain modeling such as commonKADS, etc.

References

- [1] BSI Standards, <http://www.bsi-global.com/>
- [2] Feigenbaum, E., McCorduck, P.: *The 5^o Generation*. Addison-Wesley, Reading (1983)
- [3] Gruber, T.: *Toward Principles for the Design of Ontologies Used for Knowledge Sharing*. *International Journal of Human-Computer Studies* (1995)
- [4] Posada, J., Toro, C., Wundrak, S., Stork, A.: *Ontology Supported Semantic Simplification of Large Data Sets of Industrial Plant CAD Models for Design Review Visualization*. In: Khosla, R., Howlett, R.J., Jain, L.C. (eds.) *KES 2005. LNCS (LNAI)*, vol. 3683, pp. 184–190. Springer, Heidelberg (2005)
- [5] Leroy, D.: *Standards and Publications for Engineering - Standards Organizations*, <http://www.interfacebus.com/Standards>
- [6] Soanes-Stevenson (ed.): *Oxford Dictionary of English*. Oxford University Press, Oxford (2003)
- [7] Posada, J.: *A Methodology for the Semantic Visualization of Industrial Plant CAD Models for Virtual Reality Walkthroughs*. Phd thesis, TU-Darmstadt (2005)
- [8] Toro, C., Posada, J., Oyarzun, J., Falcon, J.: *Supporting the CAD Structural Design Process with Knowledge-based Tools*. *Cybernetics and Systems* 38(5), 575–586 (2007)

A Knowledge Based System for Minimum Rectangle Nesting

Grzegorz Chmaj, Iwona Pozniak-Koszalka, and Andrzej Kasprzak

Dept. of Systems and Computer Networks, Faculty of Electronics,
Wroclaw University of Technology, 50-370 Wroclaw, Poland
iwona.pozniak-koszalka@pwr.wroc.pl

Abstract. Nesting algorithms deal with the optimal placement of shapes in specified regions subject to specified constraints. In this paper, a complex algorithm for solving two-dimensional nesting problem is proposed. Arbitrary geometric shapes are first quantized into a binary form. These binary representations are subsequently processed by operators which nest the shapes in a rectangle of minimum area. After nesting is completed, the binary representations are converted back to the original geometric form. Investigations have shown that the nesting effect is driven by quantization accuracy. Therefore, better accuracy is possible given more computing time. However, the proposed knowledge based system can significantly reduce the time of nesting, by intelligently pairing shapes, based on prior knowledge of their form.

Keywords: nesting, knowledge based system, binary quantization.

1 Introduction

The term nesting has been used to describe a wide variety of problems regarding the non-overlapping placement of a set of shapes onto a target domain, typically a surface (see an example in Fig. 1).

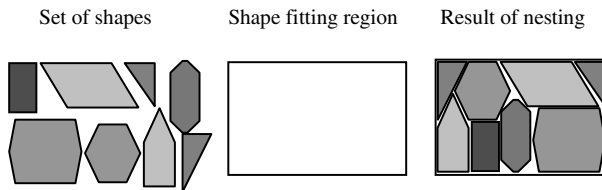


Fig. 1. A classic example of nesting

There are many classic nesting problems, including: (i) the *knapsack* problem (consider a set of shapes, and all its subsets, to cover a fixed region optimally); (ii) the *bin packing* problem (given a set of shapes and a set of bins, find the least number of bins required to pack the shapes), and; (iii) the *strip packing* problem (given a set of shapes and a strip of fixed width, minimize a length such that all the shapes can lie in the strip). This was also considered by the authors in [1, 2].

In this paper, we focus on the following nesting problem: *given*: a set of shapes and a region of infinite height and width, *find* the placement of all shapes in the region, *such that*: (i) the area of the rectangle in which all shapes are to be placed (target domain) is minimized, and (ii) the computing time for nesting is minimized.

Several approaches have been proposed for the resolution of nesting problems. Solution ideas range from simple heuristics to intricate optimization techniques, including geometric approaches [3], local search [4], ant algorithms [5], simulated annealing [6, 7], tabu search [5], and genetic algorithms [8]. However, even if the shape set is relatively small, the problem still remains computationally hard. This therefore motivated the design of an algorithm to reduce the computational time of nesting.

The main subject of this paper, namely the *QKBMR* (*Quantization with Knowledge Base in Minimum Rectangle*) nesting system, is essentially a sequential application of several algorithms performing: (i) shape conversion, by quantization, into binary format, (ii) shape pairing (applying the Min-Rectangle concept [9]), using standardization and rotation operator, and; (iii) knowledge-base exploitation. The system enables online creation of a knowledge base, (KB), which can significantly reduce the computational time needed for optimal shape fitting. The authors have already shown the remarkable improvement in nesting efficiency gained from exploiting quantization, rotation and the knowledge base. Moreover, QKBMR users can control, by judicious adjustment of the quantization parameter, the trade-off between the total time of nesting and efficient shape placement.

The rest of the paper is organized as follows: Section 2 introduces the nomenclature used throughout the paper, Section 3 describes the *QKBMR* system and Section 4 gives a short description of local algorithms. Experimental results for the *QKBMR* system are given in Section 5. Conclusions appear in Section 6.

2 Nesting Nomenclature

The following nomenclature (used in subsequent sections) is introduced:

Shape – a geometric shape is denoted $G = \{ e_n \in P: (x_m, y_n) \vee A: (x_{n1}, y_{n1}, x_{n2}, y_{n2}, r_n) \}$, where a point P lies at position (x, y) , and arc A is specified by two P s and a radius r .

Region – a fixed rectangular region R , having lower left vertex $P(0, 0)$ and having infinite height and width, is denoted $R = \{ P(x, y): 0 \geq x > \infty \wedge 0 \geq y > \infty \}$.

Mesh – a discretization of R , determined by a set of grid squares of size a , is denoted $M_{(a)} = \{ Q_{(a)}(x_m, y_n) \in R: 0 > x_m > \infty \wedge 0 > y_n > \infty \}$. Each square (quantum), is characterized by a discrete position (x_m, y_n) and is assigned a logical binary state, i.e. 1 or 0 (Fig. 2).

Intersection operator $INT(A, B)$ - returns a logical true/false result: if area A and area B intersect, then logical 1 is returned, otherwise $INT(A, B)$ returns logical 0.

Bounding rectangle $B_G(x, y)$ – the rectangle B_G , assigned to shape G , having the smallest dimensions (width x , height y) such that G can be placed inside B_G without intersecting the boundary of B_G . B_G is expressed as $B_G = \{ P(x_b, y_b): x_b, y_b \in R \wedge x_b \in \langle x_m, x_n \rangle \wedge y_b \in \langle y_m, y_n \rangle : (\forall P(x_p, y_p) \in G, x_m \leq x_p \wedge x_n \geq x_p \wedge y_m \leq y_p \wedge y_n \geq y_p) \}$.

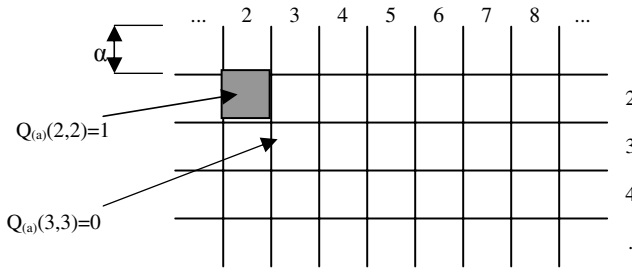


Fig. 2. Region R transformed into the $M_{(a)}$ mesh

Quantization map QM – a map converting shape G into its binary representation. The formal description is $QM(G) = (\forall Q_{(a)}(x_n, y_n) \in B, Q_{(a)}(x_n, y_n) = INT(Q_{(a)}(x_n, y_n), G))$.

Binary shape – a raster model of G , denoted $S(w, h)$, is a quantized representation of G as the result of $QM(G)$. Formally, may be described in the form: $S(w, h) = \{Q_{(a)}(x_n, y_n) \in B_G; Q_{(a)}(x_n, y_n) = INT(Q_{(a)}(x_n, y_n), G) \wedge x_n \in \langle 0, w \rangle, y_n \in \langle 0, h \rangle; B_G(x_b, y_b): x_b = a * w, y_b = a * h\}$.

Example: Binary shape in Fig. 3 is described by bit sequence $S(4,3) = 1111 0111 0011$.

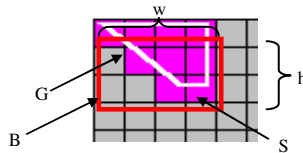


Fig. 3. Shape G and its quantized representation S

Binary Shape Set BSS – a finite set of binary shapes. $BSS = \{S_1, S_2, \dots, S_{BSSC}\}$.

AND(S_1, S_2) – binary AND operator for bit sequences S_1 and S_2 .

AND(S_1, S_2, \dots, S_n) – binary AND operation for a bit sequences S_1, S_2, \dots, S_n

OR(S_1, S_2) – binary OR operator for bit sequences S_1 and S_2 .

XOR(S_1, S_2) – binary XOR operator for bit sequences S_1 and S_2 .

NEG(S) – binary complement operator for bit sequence S , permutes 1s with 0s.

ROR1(S) – right shift operator on an ordered sequence of binary shapes: $ROR1(S_1) = S_2: S[i] = S[i+1], S[n] = S[1]; i \in \langle 2, n \rangle$, where S_1 is S before ROR1, S_2 is S after ROR1, i is a position in a bit sequence, n is the number of bits in a sequence.

ROR(S, t) – ROR1 performed t -times on S .

ROL1(S) – binary left shift operator of bit sequence by one position: $ROL1(S_1) = S_2: S[i] = S[i-1], S[1] = S[n]; i \in \langle 2, n \rangle$, where S_1 is S before ROL1, S_2 is S after ROL1,

i and n are defined as for ROR1.

ROL(S, t) – ROL1 performed t -times on S .

Best Fit $BF(S_1, S_2)$ – binary representation of the best combination of the pair of S_1 and S_2 shapes. The best fit can be denoted as $BF(S_1, S_2) = C_{12}$.

Knowledge Base Element $KBE(S_1, S_2)$ - data set with information about two given binary shapes S_1 and S_2 and their best fit combination C_{12} .

Knowledge Base KB – a knowledge base containing information about best fits for pairs of shapes, i.e. $KB = \{KBE_1, KBE_2, \dots, KBE_{KBC}\}$, where the total number of elements KBC may increase during the nesting process.

Remark. In this paper, we make the following assumptions:

- (i) a set of shapes to be nested is known before the nesting is started,
- (ii) all shapes are quantized with the same quantum size a ,
- (iii) all shapes can be rotated by an angle $gamma(k) = (k \pi) / 2, k \in \{0, 1, 2, 3\}$.

3 The QKBMR System

A schematic diagram of the *QKBMR* system is presented in Fig. 4. As mentioned in the Introduction, the general idea of the *QKBMR* system is to firstly represent the shapes in binary form (by quantization). The binary shapes are subsequently nested by operators (rotation, pairing, knowledge acquisition). Finally, using the knowledge base, shapes are fitted optimally onto the target surface.

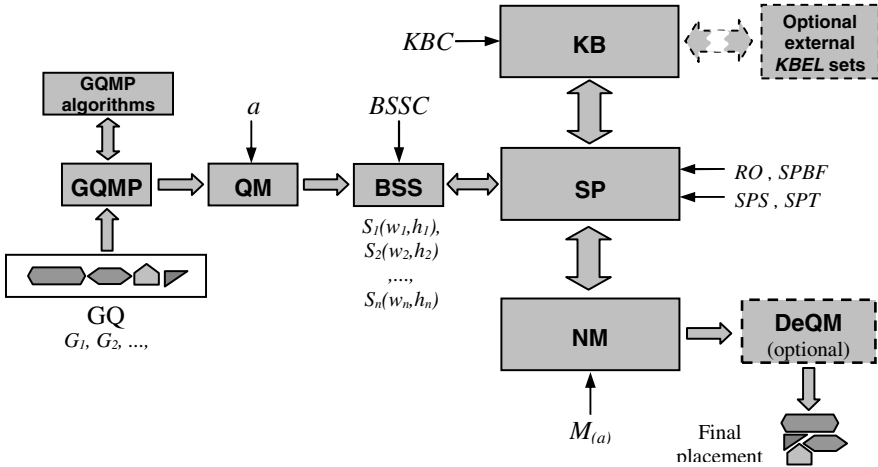


Fig. 4. Schematic diagram of the QKBMR nesting system

The components of the system are:

GQ – an input queue with a known set of shapes to be nested (the shape set can be arbitrarily chosen or generated at random),

GQMP – a module dedicated to processing input GQ before quantization. Shapes may be sorted e.g. grouped into categories (squares, polygons, ovals, etc.). Users may select any algorithm from the GQMP module for preprocessing input GQ .

QM – quantization module responsible for converting shapes into binary representations. Quantum size a determines the precision of the conversion process.

BSS – the resulting set of the $BSSC$ number of binary shapes,

SP – the shape-pairing module finds the best fit between two considered shapes. It consists of three key components: (i) shape-pair standardization (SPS), (ii) shape-pair best fit (SPBF), (iii) shape-pair truncation (SPT). Moreover, the shape rotation RO with operators (ROR1, ROL1, ... etc.) can be invoked by the SP module. This module exchanges data with the KB module e.g. by sending information about best fit data

KB – knowledge base module with information about $KBEs$ collected in the past, collected during recent nesting and possible imported from other $KBELs$ outside.

NM – nesting module responsible for the final placement of the paired binary shapes within the mesh $M_{(a)}$.

DeQM – de-quantization module converts binary shapes S back into the original Cartesian format. It uses memorized relation between the quantized form and the original form, thus, regardless of quantum size a parameter, an original shape can be located on canvas.

4 Algorithms

Quantization Map (QM) is performed for each G in GQ separately. The size of S and accuracy of QM depend on the quantum size a , The time needed to perform QM on G can be expressed as $T_{QM} = qwha$, where: w is width of B_G , h is height of B_G , q is a hardware-dependent coefficient used in QM . The QM algorithm works as follows:

1. Find $B_G(x_m, y_m)$. 2. Partition $B_G(x_m, y_m)$ into squares of size a denoted $Q_{(a)}(x_c, y_d)$ $x_c=1, 2, \dots, w$, $y_d=1, 2, \dots, h$ to obtain the mesh of the size $(w \times h)$, i.e. a 'matrix' with w columns and h rows. 3. Assign states to all $Q_{(a)}(x_c, y_d)$ elements: $\forall Q_{(a)}(x_c, y_d) \in B_G$, $Q_{(a)}(x_c, y_d) = INT(Q_{(a)}(x_c, y_d), G)$, $c \in <0, w)$, $d \in <0, h)$.

Shape-Pair Standardization (SPS) is a necessary preprocessing step to shape pairing. Two binary shapes S_1 and S_2 are standardized so that they have the same bit size. However, SPS on the shape pair (S_1, S_2) is not a commutative operation. Table 1 gives an insight into the nature of SPS: the first shape S_1 is 'centered' in its standardized format, while S_2 is initially placed in the 'top left' of its standardized format. The SPS works as:

1. For two binary shapes: $S_1(x_1, y_1)$ and $S_2(x_2, y_2)$ compute: $mx=x_1+2*y_2$, and $my=y_1+2*y_2$. 2. Standardize S_1 : (2.1) Add the number of $c=mx-x_1$ zeroes columns to S_1 . (2.2) $x_1=mx$. (2.3) Add the number of $r=mx-x_1$ zeroes rows to S_1 . (2.4) $y_1=my$. (2.5) $t=c/2+r/2*y_1$; Perform $ROR(S, t)$. 3. Standardize S_2 : (3.1) Add the number of $c=mx-x_2$ zeroes columns to S_2 . (3.2) $x_2=mx$. (3.3) Add $r=mx-x_2$ zeroes rows to S_2 . (3.4) $y_2=my$.

Table 1. An example of binary shapes S_1 and S_2 standardized with *SPS*

Before <i>SPS</i>		After <i>SPS</i>	
S_1	S_2	S_1	S_2
1111	11	00000000	11000000
0111	01	00000000	01000000
0011		00111100	00000000
		00011100	00000000
		00001100	00000000
		00000000	00000000
		00000000	00000000

Shape-Pair Best Fit (SPBF) finds the best fit of the standardized S_2 to a given standardized shape (here denoted S_1), where S_2 can be any unpaired shape from *BSS*. The algorithm compares efficiencies of the checked fits and selects the best fit. The efficiency measure denoted *EFF*, for a given fit S is defined in equation (1) below:

$$EFF = 1 / ((w(S) + 1) * (h(S) + 1)) \tag{1}$$

where $w(S)$, $h(S)$ are width and height of the fit, respectively. Here is the algorithm:

1. Shape S_1 is chosen and 'fixed'. 2. Set $GLEFF = 0$. 3. For every unpaired shape, denoted S_2 , belonging to *BSS* do: (3.1) Standardize $S_1(x_1, y_1)$ with $S_2(x_2, y_2)$. (3.2) $S_R = ROR1(S_2)$. (3.3) If $S_2[1] = 0 \vee S_R[1] = 1$, go to 3.7. (3.4) If $AND(S_2, C) = 0 \vee AND(S_R, C) \neq 0$, where $C = C_{x_2+1}$, $C_k = C_{k+1} * 2^{x_1+1} + C_0$, $C_0 = 2^{x_1}$, perform $ROR(S_2, x_2+1)$, else $ROR(S_2, 1)$. (3.5) $M = XOR(XOR(OR(S_1, S_2), S_1), S_2)$. (3.6) If $M \neq 0$, go to 3.2. (3.7) Truncate $S = OR(S_1, S_2)$. (3.8) Compute *EFF* for S . (3.9) If $GLEFF < EFF$, $GLEFF = EFF$ and $BF = S$.

Shape-Pair Truncation (SPT) the best fit S is trimmed of all rows and columns that consist only of zeros. The *SPT* is performed at the step 3.7 in the above *SP* algorithm.

Knowledge Base. *SP* algorithm communicates with the *KB*, continuously receiving and sending data. *SP* can use *KB* to optimize the search process for the best fit. Before calling the shape pairing algorithm, *SP* can ask *KB* for a specific pair S_1, S_2 . If *KB* has such a record, then it replies to *SP* with the best fit C_{12} , else it replies to *SP* with a "no result" message. In the latter case, *SP* then applies the *SP* algorithm to S_1, S_2 . The resulting best fit for these shapes is then sent to *KB* and stored.

Rotation Operator RO. rotates any shape G , around a chosen shape. *RO* may be called (i) by the shape pairing algorithm, or (ii) by the nesting module. As was mentioned in Section 2, shapes can be rotated into four possible positions. If *KB* is switched on, and *RO* is called then *SP* works as follows:

1. Shape S_1 is chosen and 'fixed'. **2.** Set $GLEFF=0; k=1$. **3.** For every unpaired shape in BSS , denoted S_2 , do: (3.1) Ask KB for information about the pair S_1 and S_2 . If KB provides a best fit BF , count EFF and update $GLEFF$, go to 5. (3.2) $k=k+1$. Rotate S_2 by the angle $gamma(k)$. (3.3) Standardize $S_1(x_1, y_1)$ with $S_2(x_2, y_2)$. (3.4) $S_R=RORI(S_2)$. (3.5) If $S_2[I]=0 \vee S_R[I]=1$, go to 3.11. (3.6) If $AND(S_2, C)=0 \vee AND(S_R, C) \neq 0$, where $C=C_{x2+1}$, $C_k=C_{k-1} * 2^{x1+1} + C_0$, $C_0=2^{x1}$, perform $ROR(S_2, x_2+1)$, else perform $ROR(S_2, 1)$. (3.7) $M=XOR(XOR(OR(S_1, S_2), S_1), S_2)$. (3.8) If $M \neq 0$, go to 3.4. (3.9) Truncate $S=OR(S_1, S_2)$. (3.10) Compute $EFF(OR(S_1, S_2))$. (3.11) If $GLEFF < EFF$, $GLEFF=EFF$ and $BF=OR(S_1, S_2)$. (3.12) If $k < 3$ go to 3.2. **4.** Send information about $BF(S_1, S_2)$, to KB . **5.** S_1 and S_2 are paired.

Example. Consider the shapes S_1 and S_2 obtained from Table 1. In Table 2, the results of shape pairing for S_1 and S_2 for three cases are presented: (i) S' - after using simple joint operation without SP , (ii) S'' - after using SP but without RO , (iii) S''' - after using SP with RO . The corresponding efficiencies (see equation (1)) are: $EFF(S')=0,033$, $EFF(S'')=0,042$, $EFF(S''')=0,050$. Shape S''' has the highest efficiency, so there are clearly remarkable gains to be made if the RO is applied.

Table 2. Binary representations of results of shape pairing (after truncation)

$S'=S(4,5)$	$S''=S(5,3)$	$S'''=S'(4,3)$
1100	01111	1111
0100	11111	1111
1111	01011	1111
0111		
0011		

Nesting Module. The nesting module decides the final placement of binary shapes within the mesh $M_{(a)}$. After receiving the set of best fit shape pairs (from the SP module), the NM exploits the task allocation algorithms (for mesh structure) described earlier by the authors in [10] for final placement on the mesh.

The nesting module can exploit shape pairing operations recursively. For example, consider the following different options:

- (a) **initial** shapes \rightarrow pairs of shapes \rightarrow **final** nesting
- (b) **initial** shapes \rightarrow pairs of shapes \rightarrow (pairs of pairs of shapes) \rightarrow ...
 \rightarrow (pairs of pairs of pairsof shapes) \rightarrow **final** nesting.

In practice, the nesting module needs a criterion to terminate excessive recursion, of the shape pairing operators. The authors are currently addressing this issue.

5 Test Results

The investigations were conducted according to the methodology described in [11] (designed by the authors). The objective was to evaluate the efficiency of the proposed mechanisms by implementing the $QKBMR$ system. Some test results are:

Experiment #1. Impact of the quantum size a on QM efficiency measured by T_{QM}
 It may be observed in Fig. 5 that this relationship may be modeled as exponential.

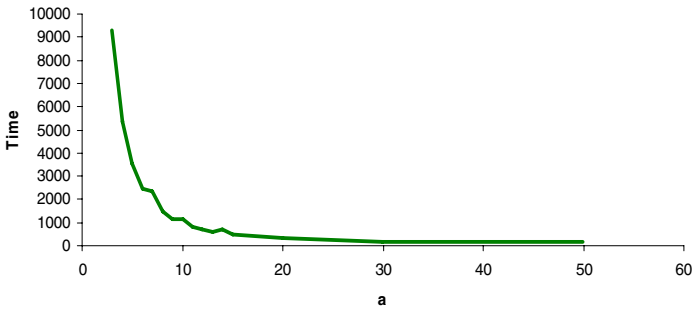


Fig. 5. Quantum size in relation to time of quantization

Experiment #2. Impact of KB on the nesting efficiency.
 In Fig. 6 an example of time performance to the successive pairing for the number of 20 shapes is presented. The remarkable advantage of applying KB may be observed.

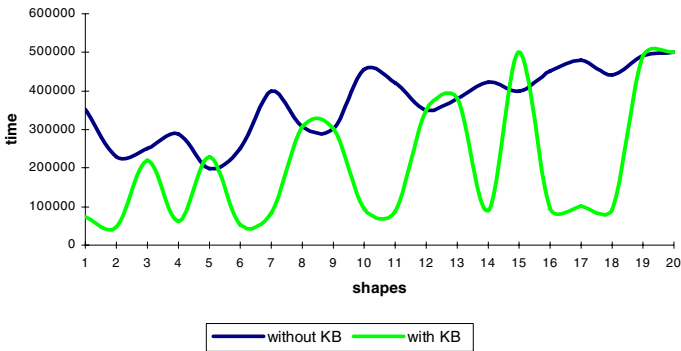


Fig. 6. Influence of Knowledge Base module on time of SP

Experiment #3. Impact of RO on the nesting efficiency.
 The efficiency was calculated with (1) for nesting made with RM turned off and RM turned on for the same data sets. The obtained values were such that the averaged $EFF (RM \text{ off}) = 0.021$ and the averaged $EFF (RM \text{ on}) = 0.038$, what may justify the statement: “ RM can increase nesting efficiency”.

6 Conclusions

The $QKBM$ R system described in this paper can significantly improve the efficiency of the nesting process.

Applying the proposed quantization algorithm in conjunction with the rotation operator ensures more accurate shape placement. A drawback of this is an increase in the computational time of the nesting process (which is strongly dependent on the chosen quantum size). However, this can be remedied using the proposed knowledge base, which can significantly reduce the computational time of the complete nesting.

The system performs well if all of the shapes belong to standard categories, because then not all shapes need to be paired: rather, best fits can be provided by exploiting the knowledge base. In practice, shapes are usually standardized and typically belong to a limited number of categories.

Future development of the *QKBMR* system will focus on creating new algorithms for the nesting module of the system, and looking for more effective shape pairing, because this module is the one with the highest computational cost.

Acknowledgments. We would like to give our appreciations to Mr. Daniel Davies for the remarkable help in preparing the final version of this paper.

References

1. Bogalinski, P., Pozniak-Koszalka, I., Koszalka, L., Kasprzak, A.: Algorithms to Solving Strip Nesting Problem based on Tabu Search and ACO. In: Proceedings to 23rd IAR Conference on Advanced Control and Diagnosis, Coventry, November 2008, pp. 240–245 (2008)
2. Chmaj, G., Koszalka, L.: Quantization Applied to 2D Shape Arrangement Problem. Acta MOSIS 112, Ostrava (2006)
3. Oliviera, F.C., Ferreira, J.A.S.: Algorithms for Nesting Problems. In: Vidal, R. (ed.). LNEMS, vol. 396, pp. 256–273. Springer, Heidelberg (1993)
4. Nielsen, B., Odgaard, A.: Fast Neighbourhood Search for the Nesting Problem. Research Report by the University of Copenhagen (2003)
5. Burke, E., Kendall, G.: Applying Ant Algorithms and the No Fit Polygon to the Nesting Problem. Scientific Report of the University of Nottingham (2002)
6. Albano, A., Sappupo, G.: Optimal Allocation of Two Dimensional Shapes using Heuristic Search Methods. IEEE Trans. on Systems, Man and Cybernetics, SMC-10, 242–248 (2001)
7. Oliveira, J., Ferreira, J.: Algorithms for Nesting. Applied Simulated Annealing (1993)
8. Rintala, T.A.: Genetic Approach to a Nesting Problem. In: Proceedings to the 2nd Nordic Workshop on Genetic Algorithms and their Applications (1996)
9. Gomes, A., Oliveira, J.: GRASP Approach to Nesting Algorithm. University of Porto (2001)
10. Koszalka, L., Kubiak, M., Pozniak-Koszalka, I.: Comparison of SBA – Family Task Allocation Algorithms for Mesh Structured Networks. In: Min, G., Di Martino, B., Yang, L.T., Guo, M., Rünger, G. (eds.) ISPA Workshops 2006. LNCS, vol. 4331, pp. 21–30. Springer, Heidelberg (2006)
11. Koszalka, L., Lisowski, D., Pozniak-Koszalka, I.: Comparison of Allocation Algorithms for Mesh Networks with Multistage Experiments. In: Gavrilova, M.L., Gervasi, O., Kumar, V., Tan, C.J.K., Taniar, D., Laganá, A., Mun, Y., Choo, H. (eds.) ICCSA 2006. LNCS, vol. 3984, pp. 58–67. Springer, Heidelberg (2006)

Evolutionary Algorithm for Solving Congestion Problem in Computer Networks

Dawid Ohia, Leszek Koszalka, and Andrzej Kasprzak

Dept. of Systems and Computer Networks, Faculty of Electronics
Wroclaw University of Technology, 50-370 Wroclaw, Poland
leszek.koszalka@pwr.wroc.pl

Abstract. The paper concerns the survivability of computer network area. The considered optimization task is the congestion problem in connection-oriented networks. This paper presents a highly configurable evolutionary algorithm together with computer experimentation system supporting its efficiency analysis. The particular emphasis is placed on parameter control and tuning process of evolutionary algorithm. The results of analysis are discussed.

Keywords: computer network, evolutionary algorithm, congestion, experiment.

1 Introduction

In connection-oriented networks, data packets sent by end-to-end communication channel are transmitted over same, established route. The examples of such technologies are Asynchronous Transfer Mode (ATM) and Multiprotocol Label Switching (MPLS). In ATM a route is called Virtual Path (VP), in MPLS it is Label Switched Path (LSP). Networks based on TCP/IP protocol stack are not connection-oriented. In such a network, a single route should be assigned for each ordered pair of nodes that require communication. When choosing routes assignment, different objectives can be considered. This paper focuses on the congestion problem [1] - an important issue in the context of network survivability i.e. an ability to return to the valid state after some kind of network failure. The considered problem can be classified as Flow Assignment (FA) problem for non-bifurcated flow (NBFA) [2]. Such optimization tasks are NP-hard [3] i.e. up till now no exact algorithms to finding a solution in polynomial time have been invented. Fortunately, in many practical applications, satisfying suboptimal solutions can be found by heuristic methods. Flow Deviation [4] is one of them, as well as its modifications [2],[3] and [5]. So-called intelligent techniques are also applied, such as simulated annealing [6] ant algorithm [7], evolutionary algorithm (EA) [3], [4], [6] and [8].

The algorithm presented in this paper, called NBFAEvol, uses solution (hence individual) representation and genetic operators applied in [8] and [9], however, NBFAEvol defines different reproduction scheme than typical evolutionary algorithm (EA) that gives opportunities to separate the process of creating offspring from the next generation selection. The act of using NBFAEvol is simultaneous searching within two solution spaces: main optimization problem (here congestion problem) search space

and EA search space itself, i.e. EA parameter setting for determining the point in EA space, from which searching the problem solution space is performed.

The main purpose of the created computer experimentation system is to enable planning and performing experiments, which examine EA search space for suboptimal parameter settings. This process is called parameter tuning phase [9]. Flexible construction of NBF AEvol allows examining many different configurations and it is open for further extension. Presented ideas are independent of considered optimization problem and relate to more general issue of applying evolutionary computation.

The reminder of the paper is organized as follows. In Section 2, the congestion problem is formulated. Section 3 describes the implemented evolutionary algorithm. Section 4 gives a review of the experimentation system. Description of performed parameter tuning phase is given in Section 5 followed by report from proper experiments. The paper concludes with a summary of the results in Section 6.

2 Problem Statement

Non-bifurcated congestion problem was defined in [1] basing on common network model that can be found e.g. in [2] and link-path notation for non-bifurcated flow [10]. Communication network is specified by a directed graph; a node in such a graph corresponds to a source, destination or point of traffic switching. Directed edge (arc) is a unidirectional link. The ordered pair of nodes that require communication is called the demand. The objective is to maximize the minimum link residual i.e. a difference between the capacity and current link load capacity (such link may be called network bottleneck). The considered problem may be formulated in the following way:

Constants. D – number of demands, E – number of arcs, V – number of nodes.

Indices: $d=1,2, \dots, D$ – demands, $e=1,2, \dots, E$ – arcs, $v=1,2, \dots, V$ – nodes, $p = 1,2, \dots, P_d$ – routes in scope of demand d .

Indexed constants. P_d – number of routes for demand d , h_d – value of demand d , c_e – capacity of link e , $\delta_{edp}=1$ (if route p for demand d contains arc e) or $\delta_{edp}=0$ (otherwise).

Variables. x_{dp} – flow allocation vector.

$$\text{Constraints.} \quad x_{dp} \in \{0, 1\} \quad \text{and} \quad \sum_{p=1}^{P_d} x_{dp} = 1 \quad (1)$$

$$f_e \leq c_e \quad \text{where} \quad f_e = \sum_{d=1}^D h_d \sum_{p=1}^{P_d} \delta_{edp} x_{dp} \quad (2)$$

Link-path notation requires a set of routes proposal for each demand, those routes are coded by constant δ_{edp} . Constraints (1) assure that for each demand exactly one route will be selected. Inequality (2) is a capacity restriction.

$$\text{Objective.} \quad \max r = \min_{e \in A} (c_e - f_e) \quad (3)$$

The objective (3) consists in finding an assignment of routes which results in network state with as broad bottleneck as possible. Non-bifurcated congestion (NBC) instance consists of two elements: (i) the Incidence Matrix which determines the structure used for graph representation, (ii) the Demand Matrix which determines demand value for each pair of nodes. The sets of routes for demands altogether is called routes database (RDB). The selection routes for RDB defines the solution space. There are two main methods for creating routes RDB: (i) K-shortest paths (KSP) where each demand is assigned by k shortest routes, with notation KSP_k , (ii) Hop limit approach (HL) where each demand is assigned by a set of all routes, which do not exceed $SP+n$ links, where SP is the shortest route between given nodes and n is the hop limit, notation HL_n is introduced. Authors examined: KSP_{10} , KSP_{30} , KSP_{100} , KSP_{1000} , and HL_2 , HL_5 .

3 NBFAEvol Algorithm

The structure of NBFAEvol, beside typical parameters as crossover probability (P_c), mutation probability (P_m) and operator selection probability [11], allows configuration of selection strategy for crossover operator, a next generation selection method, and a mechanism for preventing premature convergence of the algorithm.

Individual Representation. Representation used in [6] was applied. Single individual (solution) is represented by a vector of integer numbers. Position of an element in the vector (index) indicates the number of demand, and value at given position is the number of a selected route.

Algorithm Structure. High-level view of NBFAEvol structure is shown as below. (pseudo code notation based on Pascal language was used).

```

procedure NBFAEvol(instance, RDB, parameters)
  begin
    generation:= 1
    Create first generation
    Evaluate first generation
    while (not evolution end condition)
      begin
        Control stagnation
        Create offspring
        Evaluate offspring
        Mutate offspring
        Evaluate mutated offspring
        Select next generation
        generation := generation + 1
      end
    end
    Return best individual in the last generation
  end

```

First Generation Initiation. NBFAEvol algorithm maintains constant number of individuals in population trough whole evolution. The size of population, denoted as *pop_size*, is one of those parameters, which appear in every EA. It determines the degree of searching parallelism. Bigger population means broader search, but slower convergence. In the first generation, NBFAEvol selects routes for demands using

roulette wheel method [11]. Probability of choosing a given route is reverse proportional to its length. For NBC problem, such method results in higher average fitness value of the first generation than completely random initiation.

Reproduction. Reproduction scheme in NBFAEvol is different than in GA. The number of individuals produced in every iteration is $L = P_c * \text{pop_size}$. Reproduction is a selection of $L/2$ pairs of individuals, followed by performing crossover operation on them. The probability of selecting a given individual is proportional to its fitness value (roulette wheel method). Created offspring is mutated, an average number of mutation operations in one generation equals $P_m * L$. Offspring does not replace its parents, NBFAEvol maintains current population and the newly created offspring. Next generation selection takes place only in the last step of algorithm's iteration.

Genetic Operators. Applied operators can be divided into two categories: High-Level Operators (HLO) and Low-Level Operators (LLO). Unlike HLO, applying LLO can result in addition of new routes to RDB and in consequence in extending solution space of NBC problem. NBFAEvol implements: (i) four crossover operators: one-point crossover (IPC), multi-point crossover (MPC), uniform crossover (UC), low level crossover (LLC), and (ii) three mutation operators: one-gene mutation (1GM), multi-gene mutation (MGM), low level mutation (LLM). Operators IPC, MPC, UC, 1GM, and MGM are well known transformations described e.g. in [1], the detailed description of LLC and LLM can be found in [8].

Operators Selection Strategy. Reproduction process requires making a decision about selecting crossover operator for a given pair of individuals. NBFAEvol implements two methods: (i) constant probability strategy (ConstOp) and (ii) adaptive probability strategy (AdaptOp). In applied adaptive strategy two phases (intervals) are defined: IE (interval equal) and IA (interval adapt). IE denotes the number of generations during which all available operators are selected with equal probability. In IA phase, operators are drawn proportionally to their effectiveness in the last IE phase. Lengths of these phases are the parameters of AdaptOp. This idea is illustrated in Fig. 1.

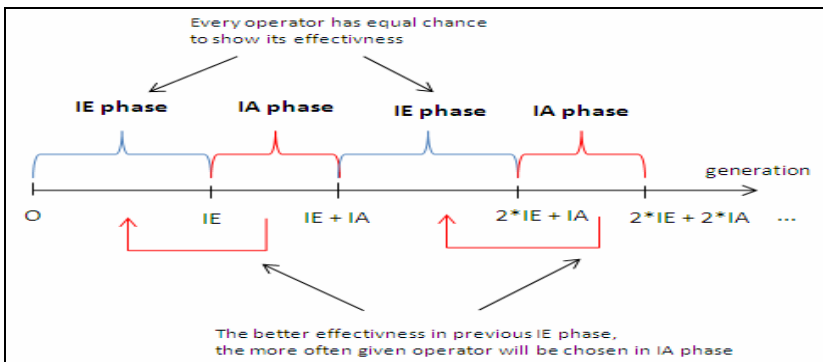


Fig. 1. Adaptive operator selection strategy

The described strategy can be classified as an adaptive method of parameter control [9]. Starting point for implementing this method was an assumption that operators can have different effectiveness measured in scores in IE phase. Scoring system looks like this: (i) 5 points if descendant is better than both parents (better fitness value), (ii) 2 points if descendant is better than one of its parents only, (iii) 0 points if descendant is worse than both of its parents. During IA phase for every operator is assigned a field in roulette with the size proportional to the total number of points collected.

Next Generations Selection. The last step in NBFAEvol algorithm is selecting individuals for the next generation. The choice is made from current population (that existed before reproduction) and newly created offspring. NBFAEvol may use one of three selection methods, including one deterministic (U+L) and two stochastic tournament strategies (T1 and T2).

Stagnation Control. It is often an end condition of the algorithm. NBFAEvol has a binary parameter, which turned on may cause attempting to break such stagnation, when it is detected. An exact mechanism is implemented: *if* for fixed number (e.g. 300) of generations all individuals have the same fitness value, *then* all except one are eliminated and new individuals take their place (initiated as the first generation). The one rescued individual spreads his genes quite quickly in a new population. Some of such attempts may give beneficial configuration of genes and improve currently best individual. In some ways, the proposed method disturbs a “natural” course of an evolution. So-called “brain storm” method [8] serves similar concept (preventing convergence) but it works in different way.

4 Experimentation System

The created experimentation system called NBFASolver consists of three modules:

NBFAExp – for creating Flow Assignment problem instances; one can use visual creator to enter network topology; also Incidence and Demand Matrices can be entered; with several modes for creating Demand Matrix: from manual to fully automated matrix generating, according to the given demand pattern,

NBFACreator - for generating routes databases,

NBFAPathsGenerator – a central part of NBFASolver for planning and executing experiments as well as presenting processed results (partial and summary).

The designed structure of the system is shown in Fig. 2.

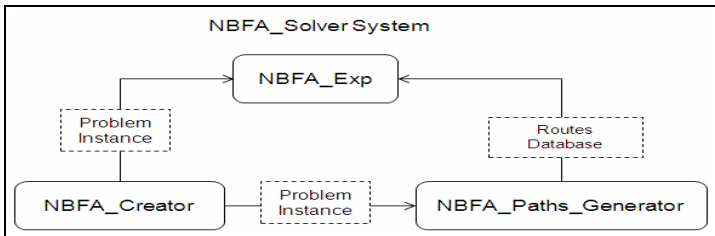


Fig. 2. Experimentation system structure

Main purpose during development of the system was to make the process of examination of the effectiveness of NBFAEvol fully automated. Nomenclature used in NBFAExp is following: *Simulation* is the smallest structural unit of experiment, it corresponds to execution of single evolution. *Sub-series* is the sequence of simulations executed with the same parameter set (necessary because of stochastic nature of EA), *Series* is the sequence of sub-series executed for subsequent parameter sets, for same instance and RDB, *Experiment* is the sequence of series. Every experiment required performing the total number of E simulation runs, where $E = \text{number of series} * \text{number of sub-series}$ in a given series * number of parameter sets. Each execution of NBFAEvol requires providing a set of seven parameters (Table 1). They establish one point in EA search space.

Table 1. Parameters of Experiment Design for EA

Symbol	Name	Available options
P_c	Crossover probability	Real number from range (0, 1>
CrossOp	Crossover operators, selection strategy	Crossover operators should be chosen as well as their selection method: ConstOpt or AdaptOpt
P_m	Mutation probability	Real number from range (0, 1>
MutOp	Mutation operators, selection strategy	Mutation operators should be chosen as well as their selection probabilities
pop_size	Population size	Integer number greater than 1
Sel (Selection)	Next generation selection strategy	- deterministic selection U + L - tournament selection T1 - tournament selection T2
Stag (Stagnation)	Stagnation control mechanism	On / Off

The experiment is carried out following the procedure as below:

```

procedure Experiment(series, parameter_sets)
begin
  x := number of series
  y := number of parameter sets
  for s := 1 to x do
    begin
      for p := 1 to y do
        begin
          z = number of subseries in series[s]
          for ss := 1 to z do
            begin
              Simulation with parameter_set[p] for series[s]
              Save result and duration of simulation
            end
          end
          Save sub-series results
        end
      end
      Save series results
    end
  end
  Process and present results of Experiment
end.

```

In a given series (for each parameter set) the number of z simulations is executed. Two results are saved: (i) an average of results in sub-series, (ii) the best result in sub-series. A comparative ratio (CR) is defined - separately for the average and the best result. These values are calculated in the following way:

$$CR_{avg} = \frac{A}{B} \cdot 100\% \quad CR_{best} = \frac{C}{D} \cdot 100\% \quad (4)$$

where A is the average of results in sub-series for a given parameter set, B is the best average result from all parameter sets, C is the best result in sub-series for given parameter set, D is the best result from the best results in sub-series from all parameter sets. Two indices for evaluating the obtained results are also applied:

- Avg CR_{avg} – an average value of CR_{avg} for a given parameter set in series,
- Avg CR_{best} – an average value of CR_{best} for a given parameter set in series.

These two values indicate which parameter set on average achieved the best results over all series. Demands are generated for randomly chosen pairs of nodes with values drawn from a given range. The following notation for instances is introduced: $N_L_R_ \%_d_X_Y$, where N - the number of nodes, L - the number of unidirectional links, R - the percentage of all pairs between which demand exists, X and Y the smallest and greatest possible value of demand (e.g.: 18_66_70%_d_40-100).

5 Investigations

The objective was: (A) to tune parameters (for matrices with homogenous demands and heterogeneous demands) i.e. finding parameter set with the greatest value of Avg CR_{avg} and parameter sets with the best value of CR_{avg} in each series, (B) to compare NBFASWP with known heuristic algorithm SWP [8] adopted for NBC problem, called here NBFASWP, and (C) to test the stagnation control.

A. Heterogeneous case. Parameter tuning phase was divided into three stages: *Stage 1* as parallel examination of parameters connected with crossover and mutation operation (without adaptive strategy), *Stage 2* as fine tuning of the first stage results, *Stage 3* as parallel comparison of: (i) the best parameter set from stage 2, (ii) adaptive operator selection strategy for different settings of IE and IA, (iii) next generation selection strategy. In experiment: pop_size=300, Sel=U+L, RDB=KSP10, End Condition=1200, Number of simulations in sub-series=3, Stag=Off. The following instances were examined: (1) 18_66_70%_d_70-200 (2) 18_82_80%_d_200-400

(3) 36_128_70%_d_40-100 (4) 36_162_80%_d_40-100.

Discussion: At the distinct stages the following observations were made: *Stage 1* – definitely the worst results were achieved when operator IPC was used; the second factor in order of importance was crossover probability, *Stage 2* – better results were achieved when beside UC, operator MPC with smaller probability was used; settings concerning mutation operator selection did not have noticeable influence on obtained results, *Stage 3* – next generation selection strategy did not have visible influence; however it is worth noticing that five best parameter sets defined tournament selection;

each configuration of IE and IA achieved comparable results as the best constant probability strategy selected after Stage 1 and Stage 2. The best Avg CR_{avg} was for:

Pc	CrossOp						Pm	MutOp		Sel
	IPC	MPC	UC	Adapt	IE	IA		1GM	MGM	
0.25	x	x	x	<input checked="" type="checkbox"/>	100	500	0.05	1.0	-	T2

A. Homogenous case. Parameter tuning phase, for selected instances, was divided into two stages: *Stage 1* - parallel examination of parameters connected with crossover and mutation operation (without adaptive strategy), *Stage 2* – parallel comparison of: (i) the best parameter set from Stage 1, (ii) the best parameter set from experiments with heterogeneous demands, (iii) adaptive operator selection strategy for different settings of IE and IA, (iv) next generation selection strategy. In experiment: pop_size=300, Sel=U+L, RDB=KSP10, End Condition=1500, Number of simulations in sub-series=3, Stag=Off. Two instances: (1) 18_66_d_100, (2) 18_82_d_100 were examined.

Discussion: It was observed that (i) significantly better results were obtained for stochastic selection of next generation, especially for T2, (ii) low effectiveness of operator LLC was observed, (iii) ConstOp strategy performs better than AdaptOp, but differences were small (a fraction of percentage), (iv) similarly as for heterogeneous demands, no influence concerning mutation operation parameters was noticed. The best Avg CR_{avg} was for two sets equally effective, as shown below:

Pc	IPC	MPC	UC	Adapt	IE	IA	Pm	1GM	MGM	LLM	Sel
0.25	0.10x	-	0.90	N	-	-	0.05	0.95	0.05	-	T2
0.25	0.10	-	0.90	N	-	-	0.03			1.0	T2

Table 3. Comparison NBFAEvol and NBFASWP

No	Instance	Evol	SWP	Profit
S 1	18_66_70%_d_70-200	3511	3228	+8.06%
S 2	18_66_80%_d_70-200	3293	2918	+11.39%
S 3	18_66_d_100	3600	3400	+5.56%
S 4	18_82_100%_d_200-400	1345	1044	+22.38%
S 5	18_82_d_100	3700	3800	-2.70%
S 6	36_128_70%_d_40-100	2103	2007	+4.56%
S 7	36_128_85%_d_40-100	1640	1497	+8.72%
S 8	36_128_100%_d_40-100	912	802	+12.06%
S 9	36_128_d_45	2345	2435	-3.84%
S 10	54_220_70%_d_50-100	150	-7	-
S 11	54_220_85%_d_50-100	1493	1471	+1.47%
S 12	72_292_50%_d_20-50	2264	2216	+2.12%

B. Comparison NBFAEvol with NBFASWP. Parameter sets for the considered instances were selected according to the following rules: (i) if given instance was tested in one of the series of parameter tuning phase, then the best parameter set for such series is used (with highest value of CR_{avg}), e.g. 18_66_70%_d_70-200, (ii) if a given instance was not tested, then the best parameter set for a given demand type and the most similar topology is used e.g. 18_66_80%_d_70-200. The results are given in Table 3, where the last column is the measure of effectiveness calculated with (5):

$$Profit = \frac{Evol - SWP}{SWP} \cdot 100\% \quad (5)$$

Discussion: It may be observed that (i) in 10 out of 12 performed simulations, better results were obtained with NBFAEvol, (ii) in simulation S10 the NBFASWP algorithm did not find feasible solution, (iii) NBFAEvol gave worse results for instances with homogeneous demand (for them only limited parameter tuning was made).

C. Stagnation control (SC) mechanism. Simulations were performed once more, this time with SC mechanism turned on. The instances were chosen including those for which NBFAEvol gave no profit (5) in B-test. Exemplary results were as follows:

Instance: 18_66_70%_d_70-200	Profit: without SC=+8.06 %, with SC=+10.73 %.
Instance: 18_82_80%_d_300-500	Profit: without SC=+22.22 %, with SC=+33.5 %.
Instance: 18_82_d_100	Profit: without SC=-2.70 %, with SC=+2.57 %.
Instance: 36_128_d_45	Profit: without SC=-3.84 %, with SC=+1.81 %.

Discussion: It was observed that better results were obtained in almost all tests performed when SC was applied (only for instance S2 Profit was worse of 0.49 %).

6 Conclusions

The created and implemented evolutionary algorithm named NBFAEvol enables examination of large EA search space. The greater effectiveness of EA may be achieved by: (i) roulette wheel method for the first population initiation, (ii) adaptive strategy of crossover operator selection, and (iii) stagnation control mechanism. The performed experiments showed an important role of parameter tuning phase. The results obtained by NBFAEvol were significantly better than those given by known NBFASWP algorithm. In the nearer future authors would concentrate in adopting ideas of multistage experiment presented in [13] to develop the experimentation system.

References

1. Przewoźniczek, M., Walkowiak, K.: Evolutionary Algorithm for Congestion Problem in Connection-Oriented Networks. In: Gervasi, O., Gavrilova, M.L., Kumar, V., Laganá, A., Lee, H.P., Mun, Y., Taniar, D., Tan, C.J.K. (eds.) ICCSA 2005. LNCS, vol. 3483, pp. 802–811. Springer, Heidelberg (2005)
2. Kasprzak, A.: Wide Area Networks. OWPW, Wrocław (1999) (in Polish)

3. Pioro, M., Deepankar, M.: *Routing, Flow, and Capacity Design in Communication and Computer Networks*. Morgan Kaufman Publishers, San Francisco (2004)
4. Fogel, D.B.: *Evolutionary Computation: The Fossil Record*. IEEE Press, Piscataway (1998)
5. Walkowiak, K.: Ant Algorithm for Flow Assignment in Connection-oriented Networks. *International Journal of Applied Mathematics and Computer Science* 15, 205–220 (2005)
6. Mahmoud, T.M.: A Genetic and Simulated Annealing Based Algorithms for Solving the Flow Assignment Problem. *International Journal of Electr. Circuits and Systems* 2 (2007)
7. Walkowiak, K.: Maximizing Residual Capacity in Connection-Oriented Networks. *Intern. J. of Applied Mathematics and Computer Science*, Article ID 72547 (2006)
8. Przewozniczek, M., Walkowiak, K.: Quasi-hierarchical Evolution Algorithm for Flow Assignment in Survivable Connection-Oriented Networks. *International Journal of Applied Mathematics and Computer Science* 16, 101–116 (2006)
9. Eiben, A.E., Hinterding, R., Michalewicz, Z.: Parameter Control in Evolutionary Algorithms. *IEEE Transactions on Evolutionary Computation* 3(2), 124–141 (1999)
10. Ford, L.R., Fulkerson, D.R.: A Suggested Computation for Maximal Multicommodity Network Flows. *Management Science*, 597–601 (1958)
11. Michalewicz, Z.: *Genetic Algorithms + Data Structure = Evolutionary Programs*. PWN, Warsaw (1999) (in Polish)
12. Ma, Q., Steenkiste, P.: On Path Selection for Traffic with Bandwidth Guarantees. In: *Proceedings of the International Conference on Network Protocols*, pp. 191–202 (1997)
13. Koszalka, L., Lisowski, D., Pozniak-Koszalka, I.: Comparison of Allocation Algorithms for Mesh Networks using Multistage Experiment. In: Gavrilova, M.L., Gervasi, O., Kumar, V., Tan, C.J.K., Taniar, D., Laganá, A., Mun, Y., Choo, H. (eds.) *ICCSA 2006*. LNCS, vol. 3984, pp. 58–67. Springer, Heidelberg (2006)

Automatic Speech-Lip Synchronization System for 3D Animation

Juan Monroy¹, Francisco Bellas¹, Richard J. Duro¹, Ruben Lopez²,
Antonio Puentes², and Jacques Isaac²

¹ Integrated Group for Engineering Research,
Universidade da Coruna, 15403, Ferrol, Spain
{jmonroy,fran,richard}@udc.es
<http://www.gii.udc.es>

² Bren Entertainment S.A., Filmax Entertainment,
Santiago de Compostela, Spain
{r.lopez,a.puentes,j.isaac}@bren.es
<http://www.bren.es>

Abstract. In the 3D animation field, the quality of productions is continuously increasing. It is a very active market with a high level of competitiveness where modest companies, in terms of budget, must reach a balance between the resources they can apply and the economic investment in a given production. Consequently, the automation of manual design processes, which are normally highly time-consuming, has become a crucial research topic for 3D animation studios. The work we are describing here presents one of these automatic tools, specifically focused on the synchronization of the speech and the lip movement of the characters, a process that is called *lipsync*. We have developed a very robust and accurate speech recognition module that together with a knowledge-based system, autonomously provides lipsync results. Additionally, the system has been integrated in the production plan of a 3D animation company, leading to drastic operator time reductions.

Keywords: 3D animation, speech recognition, knowledge-based system, lip synchronization, process automation.

1 Introduction

The synchronization of speech and lip movement, *lipsync*, implies, in its more basic form, the detection of the phonemes present in an audio file for a given language, and the moment when they happen. These phonemes must be mapped to a set of visemes (the visual reference pattern of a phoneme) that directly represent a lip configuration in the animated character. It is clear that the accuracy of speech recognition is crucial in the lipsync process.

It is a well studied problem [1][2] due to the large number of different applications in computer science that require speech-lip synchronization. In human-computer interaction systems like general virtual humans [3], mobile device interfaces [4] or assistants for hearing impaired people [5], lipsync allows using virtual characters

to interact with users in a friendly and natural way, independently of the speech source. Video games, eLearning or instant messaging are other typical application areas for lipsync. But in 3D animation productions (videos, movies, commercials, etc) is where we can find the highest level of requirements for lipsync in terms of quality [6]. Typically, the audio corresponds to an actor performance which increases the complexity of speech recognition. In addition, mainly in movies, the characters are not simple virtual humans or interfaces, but they play a role and their acting follows the director's guidelines. These two facts introduce an *artistic* component into the lipsync process for 3D animation productions that differentiates them from other applications.

The system we are presenting here starts from the need of *Bren Entertainment*, a 3D animation company, of reducing the time required to perform lipsync in their 3D productions by developing an automatic lipsync tool. The usual procedure in their movies was to perform the character's lipsync manually, due to the poor results that commercial automatic lipsync tools provide. These tools achieve accurate results with "simple" audio files, that is, when the speech is clear in terms of comprehension and when the vocabulary is typical. Such requirements seem logical, but as commented before, in 3D animation productions, the audio files correspond to an actor playing a role and this implies that the speech may be smooth, fast, exaggerated or even sung, so the audio files may be very complex to recognize. In addition, these tools did not meet the artistic requirements of the company, providing unrealistic animations, typically with excessive lip movements. Finally, commercial products are not open designs in software terms, and their integration with the 3D animation programs that studios use is not simple.

2 System Development

The first stage of this automatic lipsync system is based, as usual, on the idea of applying speech recognition to the audio files containing the phrases as spoken on the animation movie shots or scenes. This recognition allows obtaining the phonemes in these sentences and the timestamps of when the phonemes starts and finishes. Later, a phoneme-viseme assignment is carried out. A viseme, as commented before, is a visual representation for a phoneme that describes a facial and oral position for a given sound. Thus, we have a collection of timestamped visemes and we can move from the time domain into the frame domain and transform these timestamps to obtain which frames in the shot contain the assigned visemes. In the next step of the process some key frames are determined using a *knowledge-based system*. The candidate frames are those that contain visemes that provide interesting visual information to the animation. Afterwards, an animation can be constructed by assigning control values in a facial grid defined in the 3D production software to these key frames. The following sections provide a description of the whole process and the technologies it implies.

2.1 Automatic Lipsync System

To clearly understand how the system works, we have represented in Fig. 1 a diagram of the whole process. In this Figure we can see how the artist (left blocks) interacts with the system (right blocks) and the flow of information (top-down). The squares represent key functions in the whole process and the ellipses represent static pieces of information such as input or configuration parameters, provided by the artist, and the final product at the end of the process.

Phoneme recognition. The first step in the system is to apply a speech recognition tool over the input file audio provided by the artist. This can be achieved automatically when loading the shot to be animated through the 3D production software used in the company. This file usually contains the speech of a movie character for a particular shot. We use an open source speech recognition software, the CMU Sphinx-4 [7] tool from Carnegie Mellon University. In our first trials we used an operation mode for Sphinx based on a language model, but the recognition errors were too high with prototypic audio files, which, as indicated, are very complex. Consequently, we started to use an operation mode based on an acoustic model for the language and a limited language model defined as a regular grammar. This way, for a particular audio file we limit the recognition possibilities of Sphinx by providing a grammar that only matches the exact phrase spoken by the movie character. For example, if the character says “I have the donkey double parked”, we define the regular grammar with only one non-terminal symbol and several terminal symbols concatenated in the same order as pronounced:

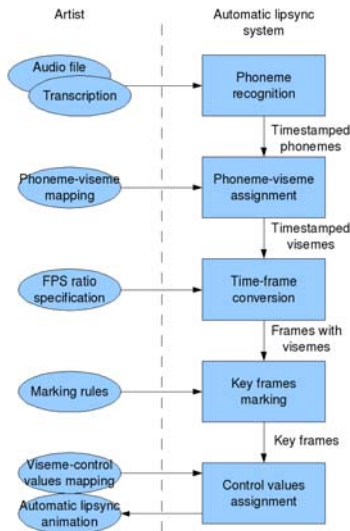


Fig. 1. Diagram of the automatic lipsync process developed

<language> = I have the donkey double parked;

So, each spoken word corresponds to a terminal symbol. This makes mandatory for the artist to provide the system with the exact transcription of the shot.

With these modifications we raise the recognition success to an admissible level (about 80%) in the easiest audio files. Due to several characteristics of the more complex audio files such as velocity, intonation or pitch, the recognition failed in some cases. To solve it, first we add some white noise to the signal to avoid silences that can "break" the recognition and, second, we slow the audio by a factor between 1.0 and 2.0. With these two processes, the recognition success tends to 100%.

At this point, the speech recognition engine does the rest, and provides as a result the ordered list of the phonemes contained in the spoken phrases (see Fig. 2). In addition, for each phoneme, Sphinx determines the timestamps of its beginning and ending, so we obtain a list of timestamped phonemes with a precision of hundredths of a second.

Phoneme-viseme assignment. The next step (see Fig. 1) depends on the viseme set defined in the 3D production software. For each subset of phonemes there is a viseme that represents it. This step is necessary because, at the animation level, the visemes play the role of the phonemes, so, the aim here is to substitute in the timestamped phoneme list by their associated viseme. The viseme set may depend on the particular production or even the character, and it is typically the artistic director the one who provides the correct phoneme-viseme mapping.

Frame domain conversion. Now that we have converted from phonemes to visemes, we have to move from the time domain to work with frames instead of seconds. Here a small conflict appears when doing the time division because two

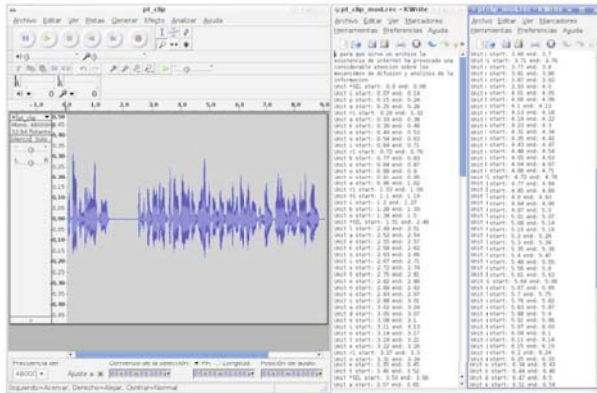


Fig. 2. Screenshot of the speech recognition result with the audio file representation (left) and the list of recognized phonemes and corresponding timestamps (right)

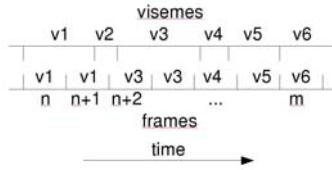


Fig. 3. Time to frame conversion

or more visemes may coexist in a frame. There are several strategies to solve this problem, but we have chosen the viseme that has a longer presence in the frame. Fig. 3 clarifies this issue, where we display the time segmented into frames (the bottom line), the visemes and their duration (the top line) and the assignment of the visemes to the frames (in between). The result of this phase is an ordered list of visemes, as many as frames has the shot.

Key frame selection. At this point, the phoneme recognition and the conversion of data has been carried out (see Fig. 1). The animation work begins here. In most 3D animation studios, the lipsync process is carried out manually, but 3D production software usually provides tools to simplify it. In this sense, the artist does not perform a frame by frame animation which, apart from being highly time consuming, from an artistic point of view would result in a poor animation in terms of the desired smoothness of the final lip movements. The typical procedure is to select some important frames (key frames) according to the visemes they contain, then configure the lip position of these visemes and finally apply an interpolation technique for the rest of the frames. The way in which the key frames are selected strongly depends on the animation guidelines provided by the artistic director. From these guidelines, we have developed a knowledge-based system (KBS) able to find and mark the key frames without the intervention of the artist. Through a phase of knowledge elicitation from multiple interviews and direct observation of an artist carrying out this task, we acquired enough knowledge to construct the KBS, that consists in a set of rules for marking key frames and a correct order to be applied.

At the end of this step, what we have is an ordered subset of the list of visemes resulting from the previous step. This subset is made up of the key frames with their associated visemes.

Control value assignment. For every key frame we have to make the correspondence between the viseme and the control values for the 3D grid that represents the lips of the movie character. Basically, this process consists in carrying out a direct assignment of control values for every point of the 3D grid. To facilitate the artist’s work, a correspondence, called *preset*, between visemes and control values exists. These presets, termed as “viseme-control value mapping” in Fig. 1, are specific to a character. This is because every character has its own and unique 3D grid. A character could have several presets to represent,

for example, different moods. The presets are provided by the artistic director depending on the shot or the scene. When the assignment is accomplished, the 3D software performs the interpolation for the non key frames and the automatic animation is considered to be finished.

As commented before, the artistic “touch” of the animators is practically impossible to achieve in an automatic way. However, our system allows the artists to fine-tune the lipsync, and they can obtain successful results with large time savings. This control value assignment represents the last step in the execution of the system as shown in Fig. 11.

2.2 3D Production Software Integration

Several steps in the lipsync process depend on or have a close relation with the 3D production software. In this case we have worked with SoftImage XSI [8]. The XSI-LSS (LipSync Server) integration has been achieved in a simple way due to the scripting and plug-in possibilities of XSI. A graphical user interface has been designed to allow the artists to select the parameters required for the automatic lipsync task, such as the audio filename or the transcription. Then, an XML file type has been created to specify the viseme-control value mapping and to represent animation, so both extremes can understand and talk the same language. This way, the integration with any other software can be easily performed.

3 Practical Application

As commented in the introduction, the origin of this work is the need of a 3D animation company to reduce resource consumption during the lipsync process in a typical audiovisual production. Currently, the automatic lipsync system is a part of the production plan in Bren Entertainment, that is, they are currently using it for new 3D productions. Fig. 4 shows a screenshot of the user interface that an artist must control to perform the automatic lipsync. In this section, we will provide some data provided obtained from the real application of the system to illustrate its behavior.

As previously indicated with regards to the speech recognition system, it was initially applied to 1050 frames of different scenes in Spanish with an 80% of success rate. But improved by means of the addition of white noise and speech speed modifications achieving 100% recognition rates even for more complex audio files.

The results provided by the whole system, including the KBS, are shown in Fig. 5, where we have represented the predicted curve (black line) compared to the curve created by an artist (dotted line) for two different controls in the 3D grid and for two different portions of the scene. These two controls are the most relevant in visual terms when moving the character lips. As we can see in the figure, the prediction is accurate enough taking into account that the differences are basically indiscernible when the curve is applied to the characters

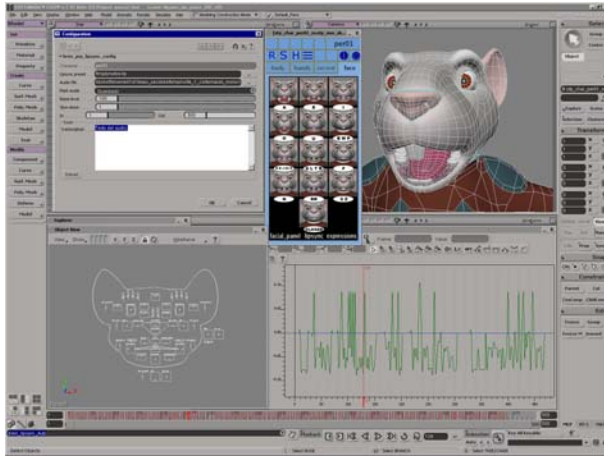


Fig. 4. Screenshot of the automatic lipsync user interface

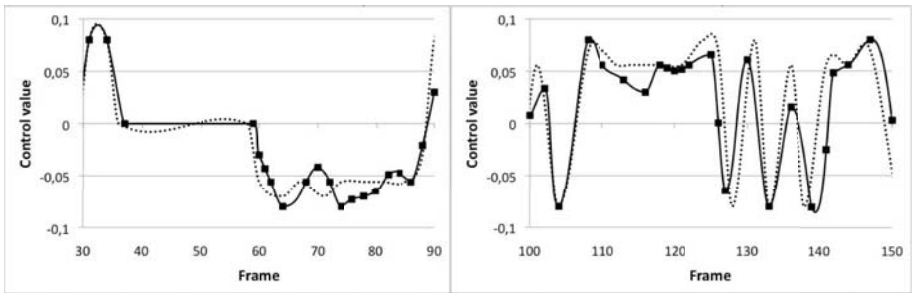


Fig. 5. Predicted control curves (black) and curves created by the artist (dotted)

grid. We must point out that the predicted curve is obtained by interpolation of the key frames (marked with black squares in the figure) that, according to the information provided by the company animators, are correctly selected in over 75% of cases. The failures occur mainly in complex audio files with expressions and moods that are artistically remarkable.

The key question is whether this version of the system reduced in fact production costs. In a current 3D production, 1050 frames of a given character have been animated in one hour with an artistic quality approved by the artistic director, using the automatic lipsync system. The same character, but in a different scene, was manually animated by an artist, and it took 24 hours (effective time) to finish 800 frames with a similar quality. This time savings imply a direct cost savings for the company which is crucial for competitiveness in the 3D animation field.

4 Conclusions

In this paper we have presented a system that performs the automatic synchronization of the speech and the lip movements of a character in a 3D animation production. The tool developed has been integrated in the production plan of an animation company and it is currently completely functional. With its application, the artists can animate a scene more than 30 times faster, which implies a very relevant productivity increase. The system uses a very robust and accurate speech recognition tool based on Sphinx-4, and which has been improved and adapted to the particular 3D animation field to obtain a near 100% recognition success after parameter tuning. In addition, the system includes a knowledge-based system that selects the key frames in the scene according to the visual relevance of the visemes they contain. This KBS reflects the basic lipsync guidelines provided by the artistic director of a given production. It has been designed in a very flexible way: it is language independent, viseme set independent and the lipsync can be adapted to the different features of the different characters that may appear in a production just by changing the viseme-control value mapping, that is, the particular 3D grid representation of a viseme for a character.

Acknowledgments. This work was partially funded by Xunta de Galicia through project PGIDIT06SIN068E.

References

1. Parke, F.I.: Computer generated animation of faces. In: Proceedings of the ACM Annual Conference, vol. 1, pp. 451–457. ACM, New York (1972)
2. Parent, R., King, S., Fujimura, O.: Issues with lip sync animation: can you read my lips? In: Proceedings of Computer Animation 2002, pp. 3–10. IEEE Press, Los Alamitos (2002)
3. Tang, S., Wee-Chung Liew, A., Yan, H.: Lip-Sync in Human Face Animation Based on Video Analysis and Spline Models. In: Proceedings of the 10th Int. Multimedia Modelling Conference (MMM 2004), pp. 1–7. IEEE Press, Los Alamitos (2004)
4. Zoric, G., Pandzic, I.S.: Automatic Lip Sync and its Use in the New Multimedia Services for Mobile Devices. In: 8th Int. Conf. on Telecommunications, pp. 353–358. IEEE Press, Los Alamitos (2005)
5. Rathinavelu, A., Thiagarajan, H., Savithri, S.R.: Evaluation of a computer aided 3D lip sync instructional model using virtual reality objects. In: Proc. 6th Intl Conf. Disability, Virtual Reality & Assoc. Tech., Esbjerg, Denmark, pp. 67–73 (2006)
6. Lewis, J.: Automated Lip-sync: Background and Techniques. *Journal of Visualization and Computer Animation* 2, 118–122 (1991)
7. The CMU Sphinx Group Open Source Speech Recognition Engines, <http://cmusphinx.sourceforge.net>
8. Softimage XSI, <http://www.softimage.com/>

Development of an Effective Travel Time Prediction Method Using Modified Moving Average Approach

Nihad Karim Chowdhury¹, Rudra Pratap Deb Nath¹, Hyunjo Lee²,
and Jaewoo Chang²

¹Department of Computer Science & Engineering, University of Chittagong, Bangladesh
nihad@cu.ac.bd, prataprudracsecu@gmail.com

²Department of Computer Engineering, Chonbuk National University, South Korea
{o2near, jwchang}@chonbuk.ac.kr

Abstract. Prediction of travel time on road network has emerged as a crucial research issue in intelligent transportation system (ITS). Travel time prediction provides information that may allow travelers to change their routes as well as departure time. To provide accurate travel time for travelers is the key challenge in this research area. In this paper, we formulate two new methods which are based on moving average can deal with this kind of challenge. In conventional moving average approach, data may lose at the beginning and end of a series. It may sometimes generate cycles or other movements that are not present in the original data. Our proposed modified method can strongly tackle those kinds of uneven presence of extreme values. We compare the proposed methods with the existing prediction methods like Switching method [10] and NBC method [11]. It is also revealed that proposed methods can reduce error significantly in compared with other existing methods.

Keywords: Intelligent transportation system, travel time prediction, moving average, NBC method, Switching method.

1 Introduction

Nowadays, travel time prediction plays an important role in ITS. Travel time prediction is becoming increasingly important with the development of the Advanced Travelers Information Systems (ATIS) [1]. Effective travel time prediction and dynamic route guidance system can assist travelers to better adjust traveler schedule [2]. Moreover, accurate prediction of travel time on road network is vital for any kinds of dynamic route guidance system. Beside this, reliable travel time information enables the generation of the shortest path from origin to destination. At the same time, time varying feature of traffic flow can extremely influence to estimate accurate travel time. Meanwhile, how to efficiently predict travel time for any road network receives a lot of attention from the researchers across the world.

In this paper, we have developed two new methods named as successive moving average and chain average for predicting reliable and accurate travel time. Moreover, this research has attempted to extend our previous travel time prediction method

which is based on Naive Bayesian Classification [11]. Both prediction methods formulate a functional relationship between traffic data as an input variables and predicted travel time as the output variable. For experimental evaluation, PNU (Pusan National University) trajectory data generator is used which provides us real trajectory data. According to experiment result, our methods exhibit satisfactory performance in terms of prediction accuracy. At the same time, the result is considered to be superior rather than other prediction methods like Switching method [10] and NBC method [11].

The remainder of this paper is organized as follows: Section 2 introduces some relevant research in this field. Characteristic of the proposed method is outlined in Section 3. A concise experimental evaluation is presented in Section 4. Finally, in Section 5, the main conclusion of this research is presented and direction of future research is outlined.

2 Literature Review and Motivation

Travel time prediction has emerged as an active and intense research area nowadays. In the literature, there are a large number of researches that can deal with accurate prediction of travel time on road networks. In the following section, a wide-ranging literature review on the topic of travel time prediction is presented.

Travel time prediction methods are broadly categorized in two parts, named as path-based estimation and link-based estimation. Most studies are focused on path travel time estimation [4] ~ [10], it is generally assumed that path travel time is the travel time between any two points in a road network. For instance, Park et al. [4] [5] proposed Artificial Neural Network (ANN) models for forecasting freeway corridor travel time rather than link travel time. One model used a Kohonen Self Organizing Feature Map (SOFM) while the other utilized a fuzzy c-means clustering technique for traffic pattern classification.

Kwon et al [6] focused on linear regression method. They used an approach to estimate travel time on freeways derived from flow and occupancy data from single loop detectors and historical travel time information. Their proposed predictor was a linear combination of the current and historical information. Zhang et al [7] proposed a method to predict freeway travel times using a linear model in which the coefficients vary as smooth functions of the departure time. A linear predictor consisting of a linear combination of the current times and the historical means of the travel times was proposed by Rice et al [9]. They presented a method to predict the time that would be needed to traverse a given time in the future.

Most recent research in this field is proposed by Erick et al [10]. They investigated a switching model which was consisted of two linear predictors. Beside this, they have shown that there is a point in future time where the linear predictor is no longer better than the historical mean. That means this point is varied according to day and time for a given roadway.

On the other hand, a few researches have investigated the use of link travel time to model travel time prediction. It is assumed that link travel time prediction is the addition of the travel times on its consisting links for a particular route. Chen et al [1] conducted a study that focused on link travel time prediction. Their study compared

the prediction accuracy under direct measurements of path-based travel time versus link-based travel times. For path-based method, probe vehicle's passing is only recorded at the beginning and the end of the path. The average probe travel time is used as the real-time observation of travel time at each time period. In link-based method, record travel times on desired links for those probes entering the links and get the average probe travel time for each link. Final travel time is calculated by adding travel times on all consisting links.

In the past research, we have observed that prediction algorithms are trained on particular route regardless of other routes in road networks. Kwon et al. [3] employed their approach for arbitrary travel routes. In their approach, at first they partition the freeway into short segments and observe future travel time on every segment. On the other hand, this approach takes more storage and computation time due to two-step computation.

In our previous work, we proposed a method by using Naïve Bayesian Classification (NBC) [11] which also focused on arbitrary travel route. The main idea of NBC method is that based on historical traffic data it will give probable velocity label for any road segment. First, user defines an origin with start time and destination. By using Naïve Bayesian classification we can find high probable velocity label for initial road segment. Then we measure end time for initial road segment and this end time becomes start time of next road segment. Finally, by adding all link travel time, we can measure approximate travel time from origin to destination.

The prediction of travel time is received an increasing attention in recent years and this motivates the development of various travel time predictors. The problem of Switching method [10] is that computational complexity arises when we measure switching point. The method developed so far all share the implicit characteristics that the route in question must be predefined. That means currently existing system usually provides prediction for only a small number of pre-determined popular routes. Since most major urban areas experience heavy congestion and an ATIS system receives lots of queries for many different routes. So it is far from satisfactory, if someone aims to build flexible ATIS that predict travel time query for arbitrary routes. Moreover, NBC method [11] predicts travel time by considering arbitrary routes. But there is a significant problem will arise when we calculate velocity level for a particular route. This route's velocity level depends on probability of time group which is divided into nine parts. As for example, if a vehicle wants to predict travel time at 6 AM then traffic information from 7 AM to 10 AM can influence that prediction. The prediction accuracy is deteriorated fast due to this inadequate time group.

To overcome the problems of the existing methods as mentioned earlier, we have proposed two new travel time prediction methods. At the same time, proposed methods are also scalable to large network with arbitrary travel routes. In the following section, we will explore the complete scenario of our proposed methods.

3 Proposed Travel Time Prediction Methods

Our proposed methods can predict travel time by analyzing the historical travel time data. As for example, a vehicle enters on a particular road segment at 10:00 AM and

wants to predict travel time. For that reason, we need to accumulate all historical travel time data for that road segment during 10:00 AM. Let $t = t_1, t_2, \dots, t_n$ be the historical travel time data for any road segment where n is the total number of historical data within a given time interval. For travel time prediction problem, we pick as our sub-problems the problem of determining the time prediction of t_i, t_{i+1}, \dots, t_j for $1 \leq i \leq j \leq n$. Let $\tau[i, j]$ be the predicted time made by computing the time t_i, t_{i+1}, \dots, t_j ; for the full problem, the predicted time to compute t_1, t_2, \dots, t_n would thus be $\tau[1, n]$. The following two methods can be used to compute $\tau[1, n]$.

3.1 Method Using Successive Moving Average

In this section, we present our new method for predicting travel time namely Successive Moving Average which can be mathematically described by following formula

$$\tau[i, j] = \begin{cases} t_i & \text{if } i = j \\ \frac{\sum_{k=i}^{j-1} \tau[i, k] + \tau[k+1, j]}{(j-i)*2} & \text{if } i < j \end{cases} \tag{1}$$

If $i = j$, then $\tau[i, j]$ is equal to t_i for $i = 1, 2, \dots, n$. In other case, if $i < j$ then we split the time sequence t_i, t_{i+1}, \dots, t_j between t_k and t_{k+1} where $i \leq k < j$. For this reason, we can compute $\tau[i, j]$ by taking the summation of sub-predicted times $\tau[i, k]$ plus $\tau[k+1, j]$ and divide that summation by $(j-i)*2$. Finally, the value of $\tau[1, n]$ indicates predicted travel time.

3.2 Method Using Chain Average

Let $t = t_1, t_2, \dots, t_n$ be the historical travel time data for any road segment within a given time interval. The value of $\tau[i, j]$ gives the predicted travel time for t_i, t_{i+1}, \dots, t_j where $1 \leq i \leq j \leq n$. Our second proposed method named as chain average, which can be written as

$$\tau[i, j] = \begin{cases} t_i & \text{if } i = j \\ \frac{\tau[i, j-1] + \tau[i+1, j]}{2} & \text{if } j > i \text{ and } i = 1 \\ \frac{\tau[i-1, j-1] + \tau[i+1, j]}{2} & \text{if } j > i \text{ and } i > 1 \end{cases} \tag{2}$$

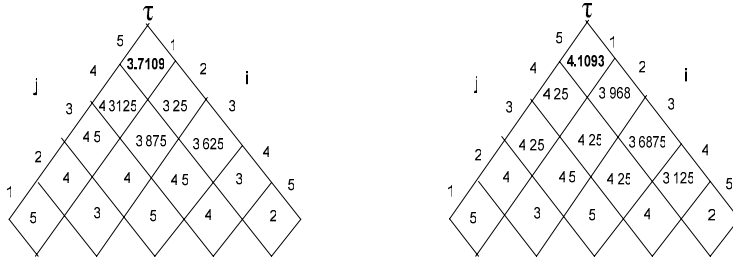
If $i = j$, then $\tau[i, j]$ is equal to t_i for $i = 1, 2, \dots, n$. Then $\tau[i, j]$ is the arithmetic mean of $\tau[i, j-1]$ and $\tau[i+1, j]$ when $i < j$ and $i = 1$. Otherwise, $\tau[i, j]$ is the arithmetic mean of $\tau[i, j-1]$ and $\tau[i+1, j]$ when $i < j$ and $i > 1$. Thus, we can compute $\tau[1, n]$ which provides us predicted travel time.

3.3 Example

Given a sequence of five sample historical travel time data within a given time interval on a particular road segment, we can predict travel time by applying our proposed methods.

Sample Historical Travel Time Data $(t_1, t_2, t_3, t_4, t_5)$: 5 sec 3 sec 5 sec 4 sec 2 sec.

Total Sample Data (n): 5



(a) Successive Moving Average Method (b) Chain Average Method

Fig. 1. τ table for proposed methods

The τ table is used for storing the value of $\tau[i, j]$. Fig.1(a) and Fig.1(b) illustrate both methods on a sample traffic data where $n=5$. In case of first method, by using equation 1 we can calculate the first value $\tau[1,2]$ as $\frac{\tau[1,1]+\tau[2,2]}{2}=4$. In this way, the value in $\tau[1,3]$ can be found by calculating arithmetic mean of $\tau[1,2]$ and $\tau[2,3]$. Furthermore, chain average method uses equation 2 for generating τ table. In chain average method, calculation of $\tau[1,2]$ is same as successive moving average method. But value in $\tau[1,3]$ depends on resultant value of $\tau[1,2]$ and value of $\tau[3,3]$. Using both layouts, the value of $\tau[1,5]$ gives us final predicted travel time. In case of successive moving average method, predicted travel time would be 3.71 sec. On the other hand, if we apply chain average method, it takes 4.10 sec to traverse that road. Therefore, predicted travel time for that road segment would be 4 sec after applying round-off operation.

4 Simulation Results

4.1 Data Set

A real data set is used in this study, which was collected by *PNU (Pusan National University)* trajectory data generator. This generator is based on real traffic situation in Pusan City, South Korea. For building *PNU* generator, they collected real traffic data by using GPS sensor. From this data, traffic pattern of Pusan city was extracted. And according to traffic pattern, generator simulates and generates trajectory data, which is

almost same as real data. The period of real traffic data covers both weekdays and weekends, and both peak hours and non-peak hours. This should adequately reflect real traffic situations. For objective and accurate evaluation of performance of the algorithms, we split data set into training and test data sets, each consisting of 365 days and 30 days. The data from 365 training days are used for fitting the model. The test data from the other 30 days are used to calculate prediction performance for all methods.

4.2 Comparison of Prediction Accuracy

To compare the accuracy between all prediction methods, we use a prediction error index, *Mean Absolute Relative Error (MARE)* [1]. As we know, *MARE* is the simplest and well-known method for measuring overall error in travel time prediction. *MARE* measures the magnitude of the relative error over the desired time range. This error measurement is defined as:

$$MARE = \frac{1}{N} \sum_t \frac{|x(t) - x^*(t)|}{x(t)} \quad (3)$$

where $x(t)$ is the observation value, $x^*(t)$ is the predicted value and N is the number of samples. In experimental evaluation, proposed methods are tested against other predictors like Switching method [10] and NBC method [11].

Relative performance between all travel time predictors is investigated in this section which is shown in Fig. 2. In this observation, prediction error of all predictors during 8AM to 6PM is examined. Our proposed two methods successive moving

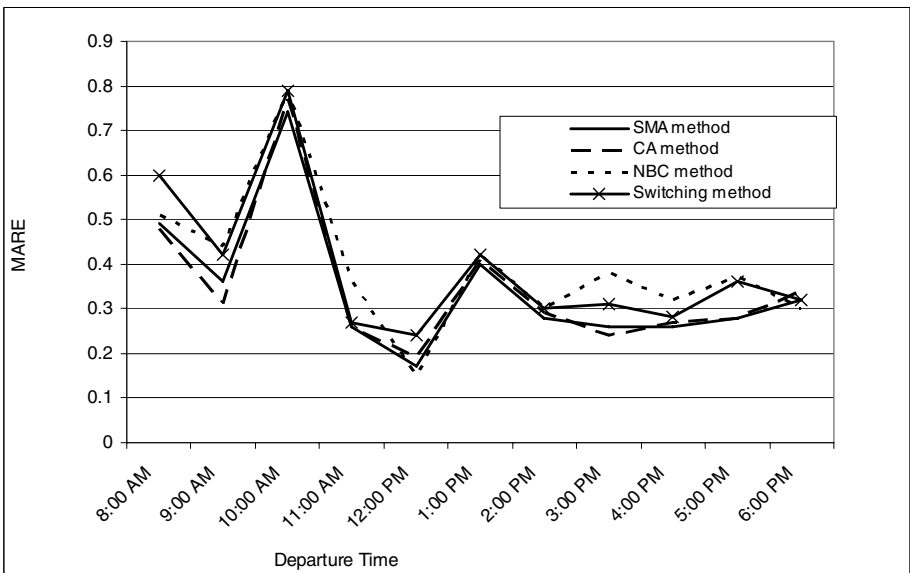


Fig. 2. MARE of each method during different time interval

average method and chain average method are denoted as SMA method and CA method respectively. We can see that, proposed methods perform much better than NBC method and Switching method. Moreover, we can note that proposed methods perform well during pick hours of a day. There are eleven test cases which are evaluated during 8AM to 6PM. In case of proposed methods, nine test cases exhibited errors less than 0.50. On the other hand, NBC method outperforms our methods in two test cases but that are not statistically significant. However, prediction error of all methods has increased dramatically at 10AM. After that, prediction error has declined significantly. At the same, our proposed methods have shown high accuracy at 12PM. In case of our methods, prediction error is varying from 0.25 to 0.40 during 2PM to 6PM. Furthermore, CA method outperforms SMA method during morning rush hour. Besides this, there are two test cases where our proposed methods provide same accuracy.

Summarized result of *MARE* for different travel time predictors are shown in Fig.3. *MARE* of successive moving average, chain average, NBC method and Switching method are 3.82, 3.84, 4.33 and 4.31 respectively. Thus, successive moving average reduces *MARE* from NBC method and Switching method by 12% and 11% respectively. A similar trend is also observed in chain average. In this case, chain average method reduces *MARE* from NBC and Switching method by 11%. Moreover, SMA method reduces *MARE* from CA method by less than 1%. So, we can say that relative performance between our proposed methods is almost same.

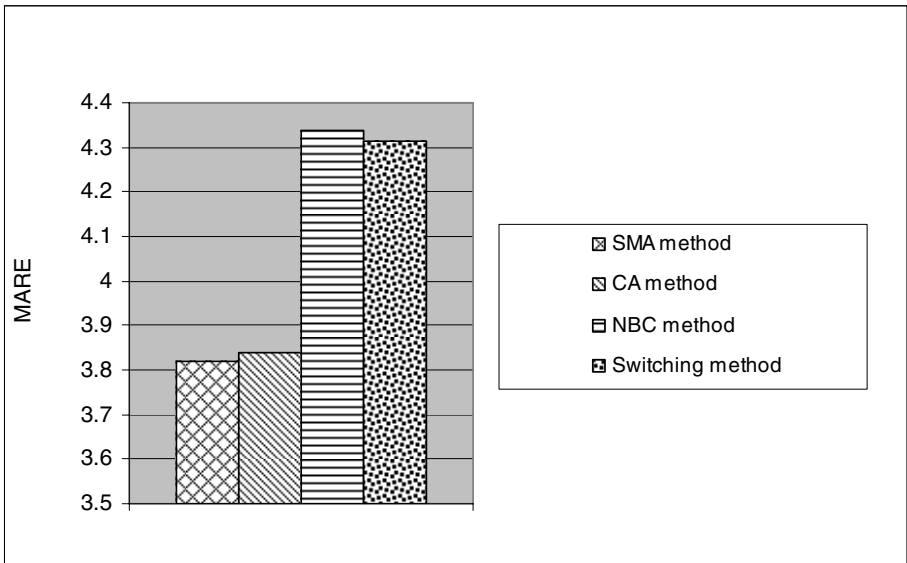


Fig. 3. Summarized MARE of each prediction method

Furthermore, we compare the effects of four methods based on some O-D (Origin-Destination) pair. The distance of each O-D pair is quite different. To reveal the relation between the distance and prediction results, Table 1 shows the distance and *MARE* from each O-D pair. In case of our methods, behavior of *MAREs* is linear and results

are varying from 0.10 to 0.25. On the other hand, *MAREs* of other methods are irregularly distributed between 0.13 and 0.20 within 15 km distance. When the distance of the O-D pair exceeds 15 km, the *MARE* of the outputs is uniformly dispersed between 0.28 and 0.38.

Table 1. Relationship between distance and prediction results (*MARE*) of some O-D pair

Method	Travel distance			
	7 km	15 km	22 km	30 km
SMA	0.10	0.16	0.22	0.23
CA	0.10	0.20	0.24	0.24
NBC	0.20	0.13	0.37	0.38
SW	0.20	0.17	0.28	0.30

5 Conclusion

This paper focuses on travel time prediction in road network for ATIS. In this paper, we have developed two methods for predicting travel time by using real traffic data from PNU trajectory generator. Compared to the other methods, simulation results suggest that proposed methods provide a more precise prediction in most test cases. Moreover, SMA method is more precise than other methods. The advantages of the proposed methods include: 1) provide an accurate prediction; 2) low cost due to simplicity. Each method needs to execute a very simple computation which reduces the complexity of the system; 3) eliminates unwanted fluctuations in the data set in comparing to conventional moving average method. Future work will include an analysis of the relationship between the length of roadways and accuracy of the prediction. Beside this, analysis of travel time prediction will be extended with respect to real field data.

Acknowledgments

This work was supported by the Korea Science and Engineering Foundation(KOSEF) grant funded by the Korea government(MEST) (No. 2009-0059417). Also we would like to thank Prof. Ki-Joune Li for providing us with the PNU (Pusan National University) trajectory data generator.

References

1. Chen, M., Chien, S.: Dynamic freeway travel time prediction using probe vehicle data: Link-based vs. Path-based. *J. of Transportation Research Record*, TRB Paper No. 01-2887, Washington, DC, pp. 157-161 (2001)
2. Chun-Hsin, W., Chia-Chen, W., Da-Chun, S., Ming-Hua, C., Jan-Ming, H.: Travel Time Prediction with Support Vector Regression. In: *IEEE Intelligent Transportation Systems Conference*, vol. 2, pp. 1438-1442 (2003)

3. Kwon, J., Petty, K.: A travel time prediction algorithm scalable to freeway networks with many nodes with arbitrary travel routes. In: Transportation Research Board 84th Annual Meeting, Washington, DC, pp. 147–153 (2005)
4. Park, D., Rilett, L.: Forecasting multiple-period freeway link travel times using modular neural networks. *J. of Transportation Research Record* 1617, 163–170 (1998)
5. Park, D., Rilett, L.: Spectral basis neural networks for real-time travel time forecasting. *J. of Transport Engineering* 125(6), 515–523 (1999)
6. Kwon, J., Coifman, B., Bickel, P.J.: Day-to-day travel time trends and travel time prediction from loop detector data. *J. of Transportation Research Record*, No. 1717, TRB, National Research Council, Washington, DC, pp. 120–129 (2000)
7. Zhang, X., Rice, J.: Short-Term Travel Time Prediction. *Transportation Research Part C* 11, 187–210 (2003)
8. Van der Voort, M., Dougherty, M., Watson, S.: Combining KOHONEN maps with ARIMA time series models to forecast traffic flow. *Transportation Research Part C* 4, 307–318 (1996)
9. Rice, J., Van Zwet, E.: A simple and effective method for predicting travel times on freeways. *IEEE Trans. Intelligent Transport Systems* 5(3), 200–207 (2004)
10. Schmitt Erick, J., Jula, H.: On the Limitations of Linear Models in Predicting Travel Times. In: *IEEE Intelligent Transportation Systems Conference*, pp. 830–835 (2007)
11. Lee, H., Chowdhury, N.K., Chang, J.: A New Travel Time Prediction Method for Intelligent Transportation Systems. In: Lovrek, I., Howlett, R.J., Jain, L.C. (eds.) *KES 2008, Part I. LNCS (LNAI)*, vol. 5177, pp. 473–483. Springer, Heidelberg (2008)
12. Han, J., Kamber, M.: *Data Mining: Concepts and techniques*, 2nd edn. Morgan Kaufmann Publishers, San Francisco (2006)

Differential Evolution and Genetic Algorithms for the Linear Ordering Problem

Václav Snášel, Pavel Krömer, and Jan Platoš

Department of Computer Science, Faculty of Electrical Engineering and Computer Science,
VŠB – Technical University of Ostrava,
17. listopadu 15, 708 33 Ostrava – Poruba, Czech Republic
{vaclav.snasel,pavel.kromer.fe,i,jan.platos.fe,i}@vsb.cz

Abstract. Linear ordering problem (LOP) is a well know NP-hard optimization problem attractive for its complexity, rich collection of testing data and variety of real world applications. It is also a popular benchmark for novel optimization and metaheuristic algorithms. In this paper, we compare the performance of genetic algorithms and differential evolution as efficient metaheuristic solvers of the LOP.

1 Introduction

The linear ordering problem is a well known NP-hard combinatorial optimization (CO) problem. It has been intensively studied and there are plenty of exact and (meta)heuristic LOP algorithms. With its large collection of well described testing datasets, the LOP represents an interesting testbed for metaheuristics aiming at the area of combinatorial optimization.

In general, a combinatorial optimization problem $\Pi = \{I, \{sol(i)\}_{i \in I}, m\}$ can be defined as a minimization or maximization problem that consists of a set of instances I for the problem, a set of feasible solutions $sol(i)$ for every instance $i \in I$ and a function $m: \{(i, q) | i \in I, q \in sol(i)\} \rightarrow \mathbb{Q}_+$, where \mathbb{Q}_+ is the set of positive rational numbers and $m(i, q)$ is the value of solution q for the problem instance i [1]. An optimal solution to an instance of a combinatorial optimization problem is such solution that has maximum (or minimum) value among all other solutions. Famous combinatorial optimization problems include among others the travelling salesman problem, the knapsack problem, and the linear ordering problem [1]. In this work, we investigate two powerful metaheuristic algorithms – genetic algorithms and differential evolution – on the linear ordering problem.

2 Linear Ordering Problem

Linear Ordering Problem (LOP) is a combinatorial optimization problem consisting in search for simultaneous permutation of rows and columns of a weight matrix C [2, 3, 4]. Consider a matrix $C^{n \times n}$, permutation Π and a cost function f :

$$f(\Pi) = \sum_{i=1}^n \sum_{j=i+1}^n C_{\Pi(i)\Pi(j)} \quad (1)$$

According to (1), LOP might be defined as search for permutation Π so that $f(\Pi)$ is maximized. Less formally, searched permutation restructures the matrix C so that the sum of its elements above main diagonal is maximized.

The LOP task is a NP-hard problem with a number of applications in scheduling (scheduling with constraints), graph theory, economy, sociology (paired comparison ranking), tournaments and archaeology among others. In economics, LOP algorithms are deployed to triangularize input-output matrices. The resulting permutation (i.e. optimal column and row ordering) provides useful information on stability of the investigated economy. The economy of a region is divided into n sectors and $n \times n$ input-output table C is constructed from the entries c_{ij} that correspond to the amount of deliveries from sector i to sector j during the investigated period [5]. In archaeology, LOP algorithms are used to process the Harris Matrix, a matrix describing most probable chronological ordering of samples found in different archaeological sites [3].

2.1 LOP Benchmarking data

There are several test collections used to benchmark LOP algorithms. They are well pre-processed, thoroughly described and the optimal (or so far best) solutions of contained LOP instances are available for comparison.

Most of investigated algorithms are tested against LOLIB library¹. LOLIB is a library of sample instances of the Linear Ordering Problem maintained at the University of Heidelberg. The library contains 50 instances of input-output matrices describing European economies in the 70s. Known optimal solutions are available. Although LOLIB features real world data, it should be considered as rather simple and easy to solve collection [6]. Anyway, it provides good benchmark for novel algorithms.

More comprehensive test collections are maintained at the University of Valencia². The UV collection includes LOLIB data as well as Stanford Graph Base (SGB), a set of more challenging (i.e. larger) input-output matrices describing US economies. Moreover, UV collection contains two sets of artificially generated LOP data and a list of known best solutions to all contained problem instances.

Mitchell and Bochers [7] published an artificial LOP data collection and LOP instance generator to evaluate their algorithm for Linear Ordering Problem. The data (from now addressed as MBLB) and code are available at Rensselaer Polytechnic Institute³.

The search space of LOP was investigated and described e.g. in [2, 3]. Some noticeable findings about known LOP benchmarking collections were published. Schiavinotto and Stützle [2, 3] showed that LOLIB and MBLB instances are significantly different, having diverse high-level characteristics of the matrix entries such as sparsity or

¹ <http://www.iwr.uni-heidelberg.de/groups/comopt/software/LOLIB/>

² <http://www.uv.es/~rmarti/paper/lop.html>

³ <http://www.rpi.edu/~mitchj/generators/linord/>

skewness. The search space analysis revealed that MBLB instances typically have higher correlation length and also a generally larger fitness-distance correlation than LOLIB instances. It suggests that MBLB instances should be easier to solve than LOLIB instances of the same dimension. A new set of large artificial LOP instances (based on LOLIB) called XLOLIB was created and made available.

In this study, the LOLIB collection was used for evaluation of performance and precision of LOP solvers based on genetic algorithms and differential evolution.

2.2 LOP Algorithms

There are several exact and heuristic algorithms for LOP problems. The exact algorithms are strongly limited by the fact that LOP is a NP-hard problem (i.e. there are no polynomial algorithms). Among the exact algorithms, branch & bound approach based on LP-relaxation of the LOP for the lower bound, a branch & cut algorithm and interior point/cutting plane algorithm attracted attention [3]. Exact algorithms are able to solve rather small general instances of LOP problem and bigger instances (the dimension of few hundred rows and columns) of certain classes of LOP [3].

A number of heuristic algorithms were used for solving LOP instances: greedy algorithm, local search algorithms, elite tabu search, scattered search and iterated local search [3]. In 2003, Huang and Lim [8] introduced hybrid genetic algorithm for LOP combining evolutionary approach with local search strategy. Recently, Snášel et al. [9, 10] applied on LOP a variant of Genetic Algorithms designed for the optimization of the turbo codes in telecommunications.

3 Genetic Algorithms

Genetic algorithms (GA) introduced by John Holland and extended by David Goldberg are wide applied and highly successful member of the wider family of evolutionary algorithms. Genetic algorithms are based on software emulation of the principles

-
- I. Define objective function
 - II. Encode initial population of possible solutions as fixed length binary strings and evaluate chromosomes in initial population using objective function
 - III. Create new population (evolutionary search for better solutions):
 - a. Select suitable chromosomes for reproduction (parents)
 - b. Apply crossover operator on parents with respect to crossover probability P_C to produce new chromosomes (offspring)
 - c. Apply mutation operator on offspring chromosomes with respect to mutation probability P_M . Add newly constituted chromosomes to new population
 - d. Until the size of new population is smaller than size of current population go back to a.
 - e. Replace current population by new population
 - IV. Evaluate current population using objective function
 - V. Check termination criteria; if not satisfied go back to III.
-

Fig. 1. A summary of genetic algorithm

of genetic evolution, Mendelian inheritance and survival of the fittest. Genetic algorithms operate on a population of artificial individuals (chromosomes) that encode potential problem solutions. Basic workflow of standard generational GA is [11]:

Genetic operators crossover and mutation are used to implement artificial evolution. Crossover is needed for varying chromosomes from one population to the next by exchanging one or more of their subparts. It mimics sexual reproduction of haploid organisms. Mutation performs random perturbation in chromosome structure. It is used for changing chromosomes randomly and introducing new genetic material into the evolving population. Genetic operators can have multiple definitions and implementations, often tailored to customize the algorithm for better performance in particular application area [11].

The termination criteria determine the end of genetic optimization. Termination criteria can include reaching global optima (often hard to detect situation), reaching limiting number of generations or certain period without progress.

3.1 Genetic Algorithms for LOP

A permutation of N symbols Π_N can be expressed as a vector $\Pi_N = (i_1, i_2, \dots, i_N)$, where $i_k \in [1, N]$ and $i_m \neq i_n$ for all $m \neq n \in [1, N]$. A sample permutation of 3 symbols ($N = 3$) is shown in (2).

$$\Pi_3 = (3, 1, 2) \quad (2)$$

Considering a data matrix $C^{3 \times 3}$, the effect of column and row permutation Π_3 is illustrated in (3).

$$C = \begin{pmatrix} 11 & 12 & 13 \\ 21 & 22 & 23 \\ 31 & 32 & 33 \end{pmatrix} \quad \Pi(C) = \begin{pmatrix} 33 & 31 & 32 \\ 13 & 11 & 12 \\ 23 & 21 & 22 \end{pmatrix} \quad (3)$$

The straightforward GA encoding of a permutation might copy the notion of Π_N and the chromosome then consists of a vector (i_1, i_2, \dots, i_N) . GA fitness function corresponds to $f(\Pi)$ from formulae (1).

The issue of genetic algorithm using above introduced permutation encoding is the implementation of crossover operator. The usage of classic crossover operators (such as one-point crossover or two-point crossover) does not guarantee validity of generated offspring chromosomes. When swapping portions of two permutation chromosomes cut at randomly selected gene, it is very likely to obtain offspring that will contain some values more than once – and thus describing binary matrix that is not valid permutation.

Better encoding of permutations for genetic algorithms can be based on random keys [12]. In random keys (RK) encoding, the chromosome consists of an array of real numbers. The sorting order of the array of numbers represents a permutation. Crossover and mutation operators can be applied on RK encoded chromosomes. An example of RK encoded chromosome and the application of one-point crossover is shown in Fig. 2.

There is one notable drawback of the RK encoding in GA. Genetic algorithms were not designed to deal with real-valued chromosomes, even though there are methods on how to process real-valued chromosomes by GA. On the other hand, there are evolutionary methods that use real values as natural encoding of problem solutions.

A chromosome with Random Key Encoding:

$$\begin{array}{l} \text{Random key: } 0.2|0.1|0.3|0.5 \\ \text{Allele: } \quad \quad 2|1|3|4 \end{array}$$

One Point Crossover of two RKE chromosomes:

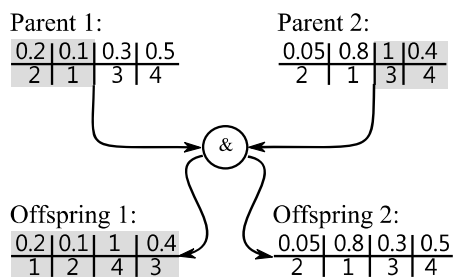


Fig. 2. Random keys encoding

4 Differential Evolution

Differential evolution (DE) is a reliable, versatile and easy to use stochastic evolutionary optimization algorithm [13]. DE is a population-based optimizer that evolves real encoded vectors representing the solutions to given problem. The real-valued nature of population vectors differentiates the DE notably from GAs that were designed to evolve solution encoded into binary or finite alphabets.

The DE starts with an initial population of N real-valued vectors. The vectors are initialized with real values either randomly or so, that they are evenly spread over the problem domain. The latter initialization leads to better results of the optimization process [13].

During the optimization, DE generates new vectors that are perturbations of existing population vectors. The algorithm perturbs vectors with the scaled difference of two randomly selected population vectors and adds the scaled random vector difference to a third randomly selected population vector to produce so called trial vector. The trial vector competes with a member of the current population with the same index. If the trial vector represents a better solution than the population vector, it takes its place in the population [13].

Differential evolution is parameterized by two parameters [13]. Scale factor $F \in (0,1+)$ controls the rate at which the population evolves and the crossover probability $C \in [0,1]$ determines the ratio of bits that are transferred to the trial vector from its opponent. The number of vectors in the population is also an important parameter of the population. The outline of DE is shown in Fig. 3.

-
- I. Initialize the population P consisting of M vectors
 - II. Evaluate an objective function ranking the vectors in the population
 - III. Create new population:
 - For $i \in \{1 \dots M\}$:
 - a. Create a trial vector $v_t^i = v_r^1 + F \cdot (v_r^2 - v_r^3)$, where $F \in [0,1]$ is a parameter and v_r^1, v_r^2, v_r^3 are three random vectors from the population P . This step is in DE called mutation.
 - b. Validate the range of coordinates of v_t^i . Optionally adjust coordinates of v_t^i so, that v_t^i is valid solution to given problem.
 - c. Perform uniform crossover. Select randomly one point (coordinate) l in v_t^i . With probability $1 - C$ let $v_t^i[m] = v^i[m]$ for each $m \in \{1, \dots, N\}$ such that $m \neq l$
 - d. Evaluate the trial vector. If the trial vector v_t^i represent a better solution than population vector v^i , replace v^i in P by v_t^i
 - IV. Check termination criteria; if not satisfied go back to III.
-

Fig. 3. A summary of differential evolution

Differential evolution is prospective method for LOP since it operates on real valued vectors and a permutation represented by RK encoding is indeed a real vector. Moreover, differential evolution has been shown to outperform genetic algorithms in some problem domains [13].

5 Algorithm Setup and Experiments

We have implemented genetic algorithms and differential evolution for LOP. The algorithms were evaluated on problem instances from the LOLIB library. The parameters of the algorithms are summarized in Table 1. They are based on best practices and initial performance experiments. Note the different interpretation of probability of crossover in genetic algorithms and differential evolution.

Table 1. A summary of algorithm parameters

Parameter	GA	DE
Population size	40	10
Terminating generation	8000	32000
Probability of crossover	$P_C = 0.8$	$C = 0.9$
Probability of mutation	$P_M = 0.02$	-
Scaling factor	-	0.9

Due to different computational costs, differential evolution was terminated after 32000 generations while genetic algorithm was terminated after 8000 generations. The execution time of both algorithms was approximately the same, about 3 seconds. To overcome the stochastic nature of both algorithms, experiments were executed several times for each problem instance and we present average lowest deviation obtained by both algorithms for each problem instance. The results of our experiments are illustrated in Figures 4 and 5 respectively.

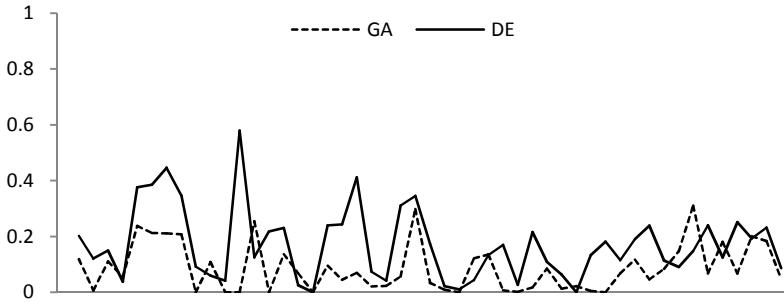


Fig. 4. Comparison of average lowest error (in %) of differential evolution and genetic algorithm for LOP. Lower is better.

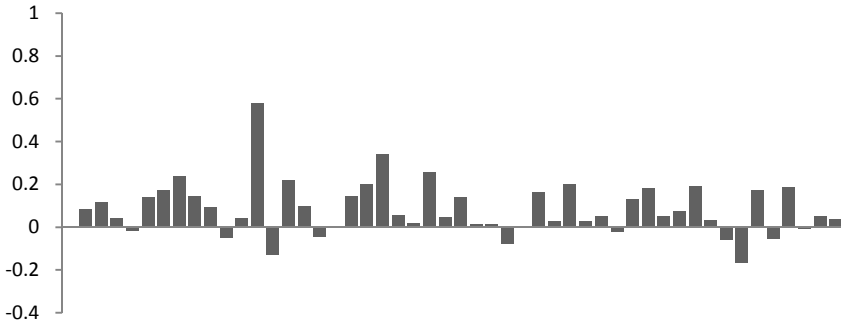


Fig. 5. Comparison of results obtained by both algorithms. Positive value denotes the loss of DE on GA and negative values correspond to loss of GA on DE in cases in which DE won.

Both algorithms reached good deviation from the optimal solution. GA featured maximum 0.31 percent deviation from optimal value; DE generated solutions delivered deviation no worse than 0.57 percent. GA delivered better solution in 39 cases while DE delivered better solution in 11 cases.

6 Conclusions

This paper investigates an implementation of differential evolution and genetic algorithms for the linear ordering problem. LOP was chosen as a representative of NP-hard combinatorial optimization problems. GA and DE represent generic optimization metaheuristics that were implemented without any additional problem specific enhancement (such as local search). Hence, they can be applied to other combinatorial optimization problems in the same form, just with different fitness function implementation. Computational experiments, performed on LOLIB library of LOP instances, showed that both algorithms have good ability to find reasonable LOP solutions with deviation from the global optimum less than 0.6 percent. GA proved its reputation of robust optimizer and delivered better solutions in 39 of 50 cases while

DE was better in 11 of 50 cases. From this point of view, DE with presented settings was outperformed as generic LOP solver by GA with random keys encoding.

It is well known that different libraries of LOP instances have fundamentally different properties. In our future work, we are going to evaluate presented implementations of GA and DE against other publicly available LOP libraries and other well known combinatorial optimization problems.

References

1. Jongen, H.T., Meer, K., Triesch, E.: *Optimization Theory*. Kluwer Academic Publishers, Dordrecht (2004)
2. Schiavinotto, T., Stützle, T.: Search space analysis of the linear ordering problem. In: Raidl, G.R., Meyer, J.-A., Middendorf, M., Cagnoni, S., Cardalda, J.J.R., Corne, D., Gottlieb, J., Guillot, A., Hart, E., Johnson, C.G., Marchiori, E. (eds.) *EvoIASP 2003, EvoWorkshops 2003, EvoSTIM 2003, EvoROB/EvoRobot 2003, EvoCOP 2003, EvoBIO 2003, and EvoMUSART 2003*. LNCS, vol. 2611, pp. 322–333. Springer, Heidelberg (2003)
3. Schiavinotto, T., Stützle, T.: *Journal of Mathematical Modelling and Algorithms* 3(4), 367–402 (2004)
4. Campos, V., Glover, F., Laguna, M., Martí, R.: *J. of Global Optimization* 21(4), 397–414 (2001)
5. Campos, V., Laguna, M., Mart, R.: 331–340 (1999)
6. Reinelt, G.: *The Linear Ordering Problem: Algorithms and Applications*. Research and Exposition in Mathematics, vol. 8. Heldermann Verlag, Berlin (1985)
7. Mitchell, J.E., Borchers, B.: Technical report, Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, NY 12180–3590. Accepted for publication in *Proceedings of HPOPT 1997, Rotterdam, The Netherlands* (September 1997)
8. Huang, G., Lim, A.: *GECCO*, pp. 1053–1064 (2003)
9. Snášel, V., Krömer, P., Platos, J.: 19th International Workshop on Database and Expert Systems Applications (DEXA Workshops 2008), Turin, Italy, September 1-5, pp. 566–570. IEEE Computer Society, Los Alamitos (2008) ISBN 978-0-7695-3299-8
10. Snasel, V., Platos, J., Kromer, P., Ouddane, N.: *Proceedings of 7th International Conference on Computer Information Systems and Industrial Management Applications (CISIM)*, Ostrava, Czech Republic, pp. 71–77. IEEE Computer Society, Los Alamitos (2008)
11. Mitchell, M.: *An Introduction to Genetic Algorithms*. MIT Press, Cambridge (1996)
12. Ashlock, D.: *Evolutionary computation for modeling and optimization*. Springer, Heidelberg (2005)
13. Price, K.V., Storn, R.M., Lampinen, J.A.: *Differential Evolution A Practical Approach to Global Optimization*. Natural Computing Series. Springer, Berlin (2005)

Determining Optimal Crop Rotations by Using Multiobjective Evolutionary Algorithms

Ruth Pavón, Ricardo Brunelli, and Christian von Lücken

Universidad Nacional de Asunción, Facultad Politécnica
Campus Universitario de la UNA, San Lorenzo, Paraguay
{rpavon, rbrunelli, clucken}@pol.una.py

Abstract. Crop rotation is a cropping system alternative that can reduce agriculture's dependence on external inputs through internal nutrient recycling. Also, it maintains long-term productivity of lands and breaks weed and disease cycles. Decision criteria to choose among competing crop rotation systems include economic and environmental considerations. Having many cultivation parcels, selection of optimal rotation alternatives may become difficult as different issues have to be analyzed simultaneously. Thus, this work proposes to use Multiobjective Evolutionary Algorithms (MOEA) to solve a multi-objective crop rotation optimization problem considering various parcels and objectives. Three outstanding MOEAs were implemented: the Strength Pareto Evolutionary Algorithm 2, the Non-dominated Sorting Genetic Algorithm and the micro-Genetic Algorithm. These MOEAS were tested using real data and their results compared using a set of metrics. The provided results have shown to be potentially useful for decision making support.

Keywords: multi-objective optimization, evolutionary algorithms, crop rotation.

1 Introduction

Crop rotation is the agricultural practice of planting different types of crops in the same land area in sequential seasons. Research and experience have proven that a well-planned crop rotation provides more consistent yields, increase profit potential, helps in pest control and maintains or improves soil structure and organic matter levels [9]. Recent literature reports attempts to model rotations and a growing interest to use computational tools to provide optimal crop rotations considering one optimization objective [6, 7, 8, 10].

The existence of contradictory objectives, as risky cultivation options with optimal returns, point out the need to analyze trade-offs solutions. It is exacerbated when analysis include other economic and environmental issues. Considering a farm with many cultivation parcels the problem may become huge. In fact, for a single season having k types of crops and m parcels the total number of possible combinations is k^m . Multi-objective Evolutionary Algorithms (MOEAs) have shown to be useful to explore large search spaces. This work proposes the use of MOEAs to determine crop rotations considering various cultivation parcels and several optimization objectives simultaneously.

This paper is organized as follows: Section 2 presents the multiobjective crop rotation problem considered in this work; Section 3 describes the use of multiobjective evolutionary algorithms to solve the problem; Section 4 presents empirical metrics considered for comparisons of three different implemented MOEAs; Section 5, presents the experimental results, and, finally, conclusions of this work are presented.

2 Multiobjective Crop Rotation Problem

This work tries to provide trade-off solutions for a crop rotation problem considering various cultivation parcels and crop alternatives with the following objectives: to minimize the total investment cost, to maximize accumulation of nutrients in soils, to maximize economic return, to minimize economic risk and to promote diversification of crops in subsequent seasons and adjacent parcels in the same season. This section briefly describes these objectives.

Total investment cost is the sum of fixed and variable costs of rotations. Fixed costs are those not influenced by chemical characteristics of soils, thus, it is the sum of fixed costs for crops in sequence. Variable cost depends of crop needs and soil characteristics. Soil test results are used to determine variable cost at beginning, estimations of soil conditions are used in next periods [2]. To obtain these costs the following data is needed: fixed costs and nutritional demands of crops; soil treatment costs for crops according soil characteristics per hectare; soil test results, size and location of parcels and information about nutrients absorbed and extracted by crops (to estimate soil conditions after a season). Total investment cost is an objective to maximize.

The amount of nutrients accumulated in each parcel is the difference between the nutrients at the beginning and the end of a rotation. The sum of the nutrients accumulated for each parcel is an objective to be maximized.

To assess future economic returns and risks historic information may be used by a scenario generator to build a scenario set S of plausible price and yield values that can be used to calculate the expected income of rotations. The total net gain is the sum of these values minus the total cost of investment. This value serves to obtain the farm return for a scenario. The mean return over S is an objective to be maximized. The standard deviation of returns can be used as a risk measure and is an objective to minimize.

Finally, crop sequences of the same family in the same plot of land are not recommended in order to control weeds and plagues. We use a penalty function that counts the occurrences of sequences with cultures of the same family during the evolutionary process and in adjacent parcels in the same period. This function is considered as an objective to be minimized.

3 Multiobjective Crop Rotation Planning Using MOEAs

In case of multi-objective problems with conflicting objectives there is not a single optimal solution but a set of solutions called the Pareto-Optimal Set. Such solutions in objective space forms the Pareto Front and represent trade-offs between the objectives

considered. A solution is said to be Pareto-optimal regarding a subset of solutions if no other solution in subset is better when all objectives are considered and preference information is not provided. A solution is true Pareto-optimal if it is non-dominated regarding the whole search space. The true Pareto-optimal set is composed of Pareto-optimal solutions. The true Pareto-optimal set and its related true Pareto-optimal front are termed as P_{true} and PF_{true} respectively [4].

To approximate the P_{true} of the problem described in previous section, this work uses an application whose architecture is composed of a scenario repository and generator, a graphic user interface, an agricultural database, and an optimization module [12]. For the optimization module, three algorithms were developed and compared: the Strength Pareto Evolutionary Algorithm 2 (SPEA2) [14], the Non-dominated Sorting Genetic Algorithm 2 (NSGA2) [5] and the micro-Genetic Algorithm (micro-GA) [3]. Due to space restrictions implemented algorithms are not presented here. A general MOEAs background is provided in [4]. However, a general view of how they work and main adjustments we done to employ MOEAs in the rotation planning problem are presented afterwards.

To solve a problem by using MOEAs a first necessary step is to encode potential solutions in such a way that these algorithms may explore the search space by means evolutionary operators. This work considers a cultivation area divided in m parcels and cultivation alternatives coded using integer indexes representing crops to cultivate [12]. In this case, 0 indicates that the parcel have not to be cultivated while numbers between 1 to 12 represent cotton, rice, oats, sugar cane, canola, rye, sunflower, corn, sesame, soybeans, sorghum and wheat respectively. A crop rotation planning is represented by a jagged array of size m , $P = [P_1, \dots, P_m]$. Each P_j contains information about a possible crop sequence in parcel i . Since crops are coded using integer indexes, P_i contains a sequence of values in $\{0, \dots, 12\}$ and $P_{i,j}$ contains the index of the crop to be cultivated in parcel i in period j . Crops may differ in sowing and harvest time as well as in the time needed to grow, this impose a restriction on the crops that can be selected in each period and makes the number of elements in each P_i differ. Figure 1 illustrates the representation of a solution.

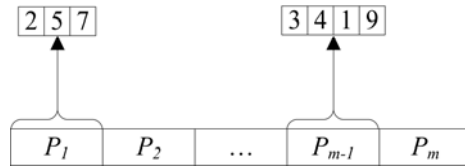


Fig. 1. Representation of a solution

After representation setting, it is necessary to integrate it to a specific MOEA algorithmic search process. This search process typically begins with the random generation of a so-called genetic initial population, i.e. a set of possible solutions of the problem or individuals. Population initialization is a crucial task in evolutionary algorithms because it can affect the convergence speed and also the quality of the final solution. In this work the initialization process is performed by random determination of plausible rotations for each parcel. Figure 2 exemplify the initialization procedure.

Each crop has a given sowing and harvest time. These periods are usually specified in month ranges, by example: sowing season for cotton is from August to September, while its harvest time is from March to May. A table with 12 rows (one for each month) containing the crops that can be cultivated in each month is used in the example. Therefore, the first step in the initialization procedure is to randomly set the initial sowing month, in case of the example it is November. Then, between crops that can be cultivated in the selected month a random pick is done; in the example, it is the crop with index 10. After the first crop is determined, the next crop in sequence is selected between crops that can be sowed in the month after the previous finishes. Taking into account regular harvesting times for crop 10, in the example, the next month is June. Again a random option for June is selected, in this case crop 5. This procedure is repeated until rotation does not exceed a given period established in advance.

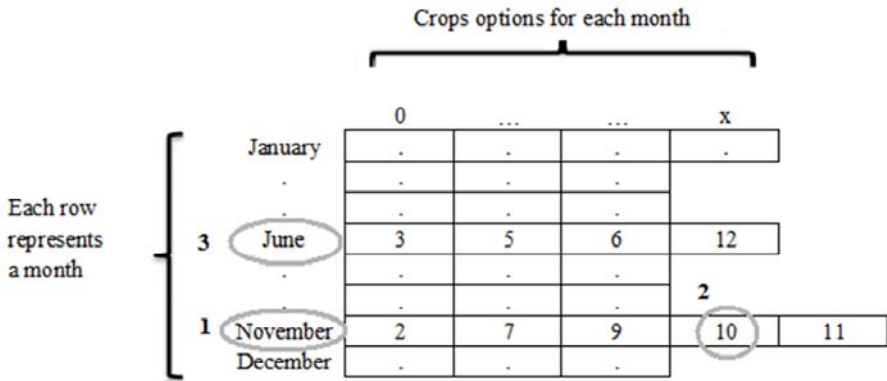


Fig. 2. Illustration of the initialization procedure

Once initialization has finished, MOEAs assigns a fitness value to each solution in population according its relative quality. Algorithms usually differ in the way fitness is assigned, but, in general, MOEAs set fitness based on Pareto Dominance concepts.

Selected individuals, called parents, interchange information by means a crossover procedure. The selection probabilities of individuals are calculated according their fitness values. Therefore, selection procedure favors to choose individuals that represent good solutions for later application of crossover operator. In this work selection is based on binary tournament selection. Crossover mechanism implemented in this work uses pairs of individuals (parents) to produce new individuals (offspring). Crossover interchanges rotations in parcels of parent solutions: a number of P_i elements are chosen and interchanged to produce new individuals. Mutation is the random modification of individuals in population. Mutation operator introduces a random walk in the search procedure. In this work, mutation is implemented by selecting a parcel and changing a crop in the sequence. If mutation produces a not viable solution a random procedure similar to the one used in initialization is applied. The overall sequence of fitness assignment, selection, crossover, and mutation is

repeated until some stop criterion is reached. This way it is expected that from the initial population a solution set evolves.

4 Performance Comparison of MOEAs

Comparing different optimization techniques experimentally always involves the notion of performance. In the case of multiobjective optimization, the definition of quality itself consists of multiple objectives [4]:

- The distance of resulting non-dominated set to the Pareto-optimal front (PF_{true}) should be minimized.
- A good, in most cases uniform, distribution of solutions found is desirable.
- The extent of the obtained non-dominated front should be maximized.

Then, to evaluate experimental results from implemented algorithms a set of metrics is used. The selected metrics are:

- *Overall true non-dominated vector generation* (OTNVG): it counts solutions in one set that are in P_{true} . For OTNVG the greater the better.
- *Error Ratio* (E): it measures the ratio between the number of solutions found by an algorithm that are in the P_{true} and the total number of solutions in it. It is expected that this value approaches zero.
- *Spacing* (S): it is an indicator of the distribution of solutions in PF_{known} and it is based on the average distance of each point from its nearest neighbor. As in case of E, the smaller the better.
- *Coverage* (C): it compares two solution sets A and B and provides information about the number of solutions in A covered by solutions in B.

Detailed explanations of the selected metrics can be found in [4, 11].

5 Experimental Results

Algorithms were tested using real soil data of a cultivable area in Alto Parana, Republic of Paraguay provided by the Soil Department of the Agronomic School of the National University of Asunción. Also, real information about prices, nutritional requirements and others, were used. Scenarios were built using the average historic yield of the area and price values were generated by a scenario generator based on normal distribution over historical data for each crop [13]. All employed data is available upon request.

For each implemented algorithm 30 different runs were carried out. For each of these runs algorithms were initialized at random with a different seed. Executions use the following parameters: number of generations is 10000; genetic population size of 100 individuals; external population size of 80 individuals; crossover probability is set to 0.8 and mutation probability is 0.1. Solutions of the 30 runs of each algorithm were grouped and non-dominated individuals extracted. This way we obtain the Pareto Front of each algorithm

The number of solutions in the final solution set of each algorithm that are obtained by individual runs is summarized in Table 1 using the minimum, maximum, average

and standard deviation of the number of solutions obtained in each run that finally becomes a solution in the Pareto Front provided by the MOEA. According to Table 1, considering non-dominated solutions provided by 30 runs, SPEA2 provides a final solution set with 75 elements, in average 9 solutions in each run are non-dominated regarding SPEA final solution set, and the minimum and maximum values are 6 and 24 respectively. From the considered algorithms NSGA2 is the one that provide the greatest solution set, and also it is the one that have the greatest standard deviation. The micro-GA shows the minimum deviation but compared to NSGA2 a reduced number of solutions.

Table 1. Solutions provided by individual runs to the final solution set of each MOEA

<i>Algorithm</i>	<i>Total</i>	<i>Min.</i>	<i>Max.</i>	<i>Average</i>	<i>Std. Dev.</i>
SPEA2	75	6	24	9	0.081
NSGA2	294	9	69	29	0.19
micro-GA	186	7	39	20	0.078

Solution sets obtained by the algorithms were evaluated over the performance metrics mentioned in Section 4. Since some of these metrics require PF_{true} to be computed, an approximation of it was calculated from the non-dominated solutions in the union set of all obtained results. This experimental set is taken as the PF_{true} of reference.

Table 2 presents the values obtained for the metrics OTNVG, E and S. From this table we show that the micro-GA obtains the smallest error ratio for the parameters used. From spacing metric is the NSGA2 the one with the best value that in fact, as previously shown in Table 1, is the one that propose more solutions. For the OTNVG metric the micro-GA obtain the best performance almost doubling the NSGA2 and ten times the value obtained by the SPEA2. SPEA2 is in fact the one with the worse performance between the three considered algorithms for all analyzed criteria.

Table 2. Values of E, S and OTNVG metrics for the different algorithms

<i>Algorithm</i>	<i>E</i>	<i>Algorithm</i>	<i>S</i>	<i>Algorithm</i>	<i>OTNVG</i>
micro-GA	0.27	NSGA2	0.017	micro-GA	45
NSGA2	0.48	micro-GA	0.019	NSGA2	28
SPEA2	0.84	SPEA2	0.101	SPEA2	4

As observed in Table 3, solutions provided by the micro-GA are better or equal (covers) than a great percentage of solutions number of solutions while their solutions are almost not covered.

Solutions provided by the system were compared with traditional rotations that are usually applied in the selected area. The whole parcel set was evaluated considering these traditional options and results compared for non-dominance with those obtained by

Table 3. Values of Coverage metric for the different algorithms

	micro-GA	NSGA2	SPEA2
micro-GA	-	0.72	0.83
NSGA2	0.13	-	0.46
SPEA2	0.05	0.18	-

the system. The result was that four traditional rotations are non-dominated regarding the set of solutions provided by the application. Figure 3 presents a bubble graphic with non-dominated solutions in objective space both those found by the proposed application and the ones representing traditional rotation alternatives as indicated. In this case mean return is on x axis, total investment cost is on axis y and bubble volume is the standard deviation of returns. As can be seen the proposed solutions ranges from low cost/low returns to solutions having a high return at a higher cost. This way, decision makers may have a better understanding of the problem at hand and choose an alternative that best fits its expectations and economic capacities. The application was evaluated for usefulness by specialist in the field.

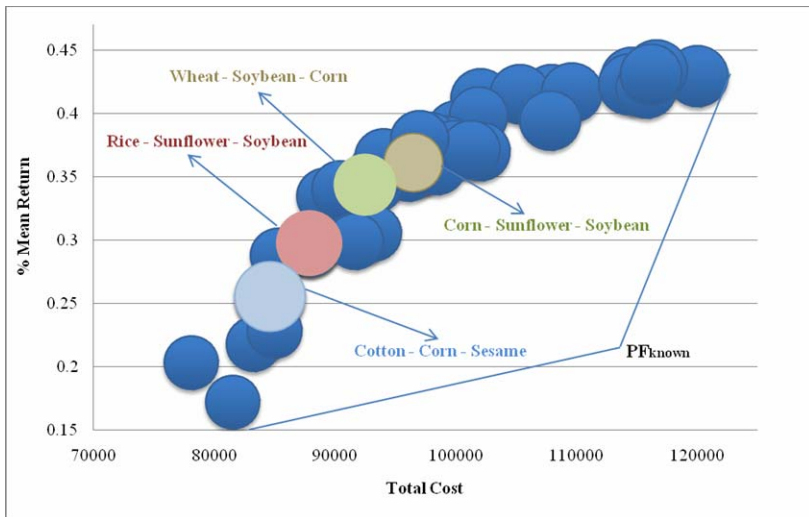


Fig. 3. Bubble graphic with solutions in objective space

6 Conclusions

This work defines a multiobjective crop rotation problem considering economic and ecologic factors simultaneously and proposes the use of an application based on MOEAs to solve it in order to provide a tool to assist decision makers in planning sustainable rotations.

Using the multi-objective framework, three multiobjective evolutionary algorithms were implemented; considering 12 types of crops, the soil analysis, the nutrient balance, the botanical family of crops, the historical yield and prices, the sowing and harvesting seasons, the fixed and variable costs of production.

The solutions obtained provide the decision-taker a broader view of the problem taking into account both goals between them in the optimization. Moreover, considering the performance tests carried out, the micro-genetic algorithm GA obtained better performance for the problem considering OTNVG, E, S and Coverage metrics.

References

1. Baker, R., Ball, S.T., Flynn, R.: Soil analysis: A key to soil nutrient management. Technical report, Cooperative Extension Service College of Agriculture and Home Economics. New Mexico State University (2002)
2. Ciampitti, I., García, F.: Requerimientos nutricionales. Absorción y extracción de macronutrientes y nutrientes secundarios: Cereales, Oleaginosos e Industriales. *Informaciones Agronómicas* 11(33), 1–4 (2007)
3. Coello Coello, C.A., Toscano, G.P.: A micro-genetic algorithm for multiobjective optimization. In: Zitzler, E., Deb, K., Thiele, L., Coello Coello, C.A., Corne, D.W. (eds.) *EMO 2001*. LNCS, vol. 1993, pp. 126–140. Springer, Heidelberg (2001)
4. Coello Coello, C.A., Van Veldhuizen, D.A., Lamont, G.B.: *Evolutionary Algorithms for Solving Multi-Objective Problems*. Kluwer Academic Publishers, Dordrecht (2002)
5. Deb, K., Pratap, A., Agarwal, S., Meyarivan, T.: A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation* 6, 182–197 (2002)
6. Detlefsen, N., Jensen, A.L.: Modelling optimal crop sequences using network flows. *Agricultural Systems* 94, 566–572
7. Dogliotti, S., Rossing, W.A.H., Van Ittersum, M.K.: ROTAT, a tool for systematically generating crop rotations. *European Journal of Agronomy* 19, 239–250 (2003)
8. El-Nazer, T., McCarl, B.A.: The choice of crop rotation: A modeling approach and case study. *American Journal of Agricultural Economics* 68(1), 127–136 (1986)
9. Food and Agriculture Organization, <http://www.fao.org>
10. Santos, L.M.R., Santos, R.H., Arenales, M.N., Raggi, L.A.: Um modelo para a programação de rotações de culturas. *Pesquisa Operacional* 27, 535–547 (2007)
11. Van Veldhuizen, D.A., Lamont, G.B.: On Measuring Multiobjective Evolutionary Algorithm Performance. In: *Congress on Evolutionary Computation*, vol. 1, pp. 204–211. IEEE Service Center, Los Alamitos (2000)
12. Von Lücken, C., Brunelli, R.: Crops Selection for Optimal Soil Planning using Multiobjective Evolutionary Algorithms. In: *Twenty-Third AAAI Conference on Artificial Intelligence*, Chicago, pp. 1751–1756 (2008)
13. Yu, L.Y., Ji, X.D., Wang, S.Y.: Stochastic programming models in financial optimization: A survey. *Advanced Modeling and Optimization* 5(1) (2003)
14. Zitzler, E., Laumanns, M., Thiele, L.: SPEA2: Improving the Performance of the Strength Pareto Evolutionary Algorithm. Technical Report, Computer Engineering and Communication Networks Lab (TIK), Swiss Federal Institute of Technology (2001)

Object Recognition by Permanence of Ratios Based Fusion and Gaussian Bayes Decision

Tuan D. Pham

ADFA School of Information Technology and Electrical Engineering
The University of New South Wales
Canberra, ACT 2600, Australia
t.pham@adfa.edu.au

Abstract. Object recognition in digital image processing is the task of finding a particular object in an image. Although there are many pattern recognition methods developed for handling the problem of object recognition, it is still a challenging task in computer vision systems and image understanding. This paper presents a new model for object recognition using the concepts of Bayes classifier, fusion of probability measures, and the permanence of ratios.

1 Introduction

An effective method for identifying an object in an image is by the best matching of its descriptive features. There are many types of features which can be extracted to provide a distinctive description of the object. This description extracted from a training image can then be used to identify the object when attempting to locate the object in a test image containing many other objects. It is important that the set of features extracted from the training image is robust to changes in image scale, noise, illumination and local geometric distortion, for performing reliable recognition.

Image analysis for object recognition by information fusion approach has been reported as a useful approach for improving recognition rates [1]-[3]. Fusion of different features for classification with large image data can be very cost-effective for practical purposes [4], [5]. There are many mathematical operators developed for data fusion such as the averaging rule, multiplication rule, probabilistic models, mathematical theory of evidence, machine learning methods, and fuzzy integral [6]. The rationale of combining knowledge from various sources is that there is always difficult or impossible to design a single classifier or to use a single feature for pattern classification to achieve the best results, because a particular classifier or feature can only be robust for handling a particular identity of an object, which may vary under different settings. Besides, different problems may require different data fusion methods to obtain effective solutions depending on the types of features.

This paper discusses a probabilistic approach for image recognition by combining evidences from multiple sources of image data where the strong assumption

of data independence is relaxed [7]. The utilization of such novel idea appears to be promising for image understanding but it is still rarely explored except the work in [8]. In order to obtain the conditional evidences in terms of probability measures to be incorporated in the fusion scheme, the Gaussian probability distribution function is utilized. Combined evidences are then used by Bayes decision rule for object classification.

The rest of this paper is organized as follows. Section 2 outlines the procedures for estimating conditional probabilities and Bayes decision rule. Section 3 describes the framework of the probabilistic information fusion based on the well-known paradigm of permanence of ratios in engineering approximation. Section 4 presents an example to illustrate the performance of the proposed method for object recognition. Finally, concluding remarks are given in Section 4.

2 Estimating Conditional Probabilities and Gaussian Bayes Decision

Pattern recognition using decision-theoretic framework is based on a discriminant or decision function to assign the unknown pattern to the best match. Let $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$ be an n -dimensional feature vector; and $\Omega = \{\omega_1, \omega_2, \dots, \omega_m\}$ the set of m distinct patterns. The Bayes classifier for a 0-1 loss function is expressed as [9]

$$d_i(\mathbf{x}) = p(\mathbf{x}|\omega_i)P(\omega_i); \quad i = 1, \dots, m. \quad (1)$$

where $d_i(\mathbf{x})$ is a decision function that measures how likely the unknown pattern \mathbf{x} belongs to the i th pattern class, $p(\mathbf{x}|\omega_i)$ is the probability density function of the feature vector of class ω_i , and $P(\omega_i)$ is the probability that class ω_i occurs.

The recognition procedure is to compute the m decision function $d_i(\mathbf{x})$, $i = 1, \dots, m$; and then assign the pattern to the class whose decision function value is maximum. Using the Gaussian probability distribution function, its n -dimensional form is given as

$$p(\mathbf{x}|\omega_i) = \frac{1}{(2\pi)^{n/2}(\det \mathbf{C}_i)^{1/2}} e^{-\frac{1}{2}[(\mathbf{x}-\mathbf{m}_i)^T \mathbf{C}_i^{-1}(\mathbf{x}-\mathbf{m}_i)]} \quad (2)$$

where \mathbf{C}_i and \mathbf{m}_i are the covariance matrix and mean vector of the pattern feature of class ω_i , and $\det \mathbf{C}_i$ is the determinant of \mathbf{C}_i .

Using the monotonically increasing property of the logarithm, the decision function $d_i(\mathbf{x})$ has the following logarithmic form

$$d_i(\mathbf{x}) = \ln[p(\mathbf{x}|\omega_i)P(\omega_i)] = \ln p(\mathbf{x}|\omega_i) + \ln P(\omega_i) \quad (3)$$

The substitution of the expression for the Gaussian probability distribution function expressed in (2) into (3) and after some mathematical rearrangement give

$$d_i(\mathbf{x}) = \ln P(\omega_i) - \frac{1}{2} \ln(\det \mathbf{C}_i) - \frac{1}{2}[(\mathbf{x} - \mathbf{m}_i)^T \mathbf{C}_i^{-1}(\mathbf{x} - \mathbf{m}_i)] \quad (4)$$

The equation expressed in (4) is known as the Bayesian decision function for Gaussian pattern class ω_i under the condition of a 0-1 loss function.

3 Information Fusion Using Permanence of Ratio Hypothesis

Based on the engineering paradigm of the permanence of updating ratios, which asserts that the rates or ratios of increments are more stable than the increments themselves, as an alternative to the assumption of the full or conditional independence of probabilistic models; Journel introduced a scheme for information fusion of diverse sources [7]. This scheme allows the combination of data events without having to assume their independence. This information fusion is described as follows.

Let $P(A)$ be the prior probability of the occurrence of data event A ; $P(A|B)$ and $P(A|C)$ be the probabilities of occurrence of event A given the knowledge of events B and C , respectively; $P(B|A)$ and $P(C|A)$ the probabilities of observing events B and C given A , respectively. Using Bayes' law, the posterior probability of A given B and C is

$$P(A|B, C) = \frac{P(A, B, C)}{P(B, C)} = \frac{P(A)P(B|A)P(C|A, B)}{P(B, C)} \quad (5)$$

The simplest way for computing the two probabilistic models is to assume the model independence, giving

$$P(C|A, B) = P(C|A) \quad (6)$$

and

$$P(B, C) = P(B)P(C) \quad (7)$$

Thus, (5) can be rewritten as

$$P(A|B, C) = \frac{P(A)P(B|A)P(C|A)}{P(B)P(C)} \quad (8)$$

or

$$\frac{P(A|B, C)}{P(A)} = \frac{P(A|B)}{P(A)} \frac{P(A|C)}{P(A)} \quad (9)$$

However, the assumption of conditional independence between the data events usually does not statistically perform well and leads to inconsistencies in many real applications [7]. Therefore, an alternative to the hypothesis of conventional data event independence should be considered. The permanence of ratios based approach allows data events B and C to be incrementally conditionally dependent and its fusion scheme gives

$$P(A|B, C) = \frac{1}{1+x} = \frac{a}{a+bc} \in [0, 1] \quad (10)$$

where

$$a = \frac{1 - P(A)}{P(A)}, \quad b = \frac{1 - P(A|B)}{P(A|B)}, \quad c = \frac{1 - P(A|C)}{P(A|C)}, \quad x = \frac{1 - P(A|B, C)}{P(A|B, C)}$$

An interpretation of the fusion expressed in (10) is as follows. Let A is the target event which is to be updated by events B and C . The term a is considered as a measure of prior uncertainty about the target event A or a distance to the occurrence of A without any updated evidence. We have $a = 0$ for $P(A) = 1$ if target event A is certain to occur; and $a = \infty$ for $P(A) = 0$ if A is an impossible event. Likewise, b and c measure the distances to A knowing about its occurrence after observing evidences given by B and C , respectively. The term x is the distance to the target event A occurring after observing evidences given by both events B and C . The ratio c/a is then the incremental (increasing or decreasing) information of C to that distance starting from the prior distance a . Similarly, the ration x/b is the incremental information of C starting from the distance b . Thus, the permanence of ratios provides the following relation

$$\frac{x}{b} \approx \frac{c}{a} \quad (11)$$

which says that the incremental information about C to the knowledge of A is the same after or before knowing B . In other words, the incremental contribution of information from C about A is independent of B . This expression relaxes the restriction of the assumption of full independence of B and C .

For the generation of k data events E_j , $j = 1, \dots, k$; the conditional probability provided by a succession of $(k - 1)$ permanence of ratios is given as

$$P(A|E_j, j = 1, \dots, k) = \frac{1}{1 + x} \in [0, 1] \quad (12)$$

where

$$x = \frac{\prod_{j=1}^k d_j}{a^{k-1}} \geq 0$$

$$a = \frac{1 - P(A)}{P(A)}$$

$$d_j = \frac{1 - P(A|E_j)}{P(A|E_j)}, \quad j = 1, \dots, k$$

It is clear that expression (12) requires only the knowledge of the prior probability $P(A)$, and the k elementary single conditional probabilities $P(A|E_j)$, $j = 1, \dots, k$, which can be independently computed.

4 Experimental Results

We used the multispectral image data described in [9] to illustrate the application of Bayes classifier and permanence of ration based fusion for object recognition.

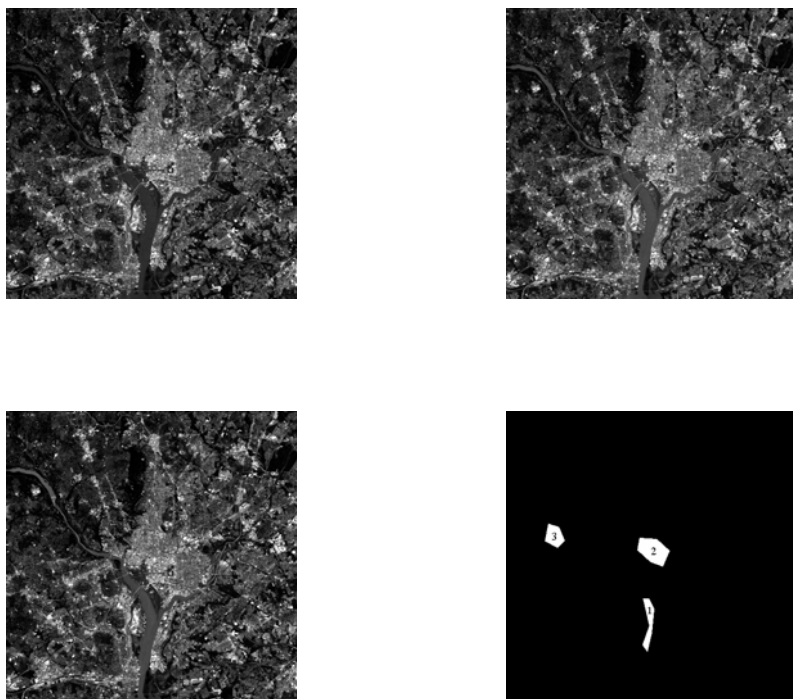


Fig. 1. Image sources in blue (top left), green (top right), and red (bottom left) visible wavelengths; and marked samples (bottom right) showing water (1), urban development (2), and vegetation (3) [9]

Figure 1 shows the three multispectral images in blue (a), green (b), and red (c) visible wavelengths; and the marked image (d) showing sample regions of water (1), urban development (2), and vegetation (3). The task is to classify these three classes of regions.

To apply the information fusion scheme expressed in (12) with $k = 3$, the estimates of prior probabilities for each image class were obtained by dividing the number of pixels of each class by the total number of pixels of all three classes. The conditional probabilities for the three classes were determined using the function defined in (2) assuming the Gaussian distribution function (GDF). The combined probabilities were then used for decision making using the GDF-based Bayes classification function defined in (3). The recognition rates obtained from the probabilistic fusion scheme are given in Table 1 which also shows the rates obtained by the multispectral fusion by forming the pattern vectors from the three images in visible wavelengths plus an infrared image [9]. Results obtained from the multispectral fusion were based on the Gaussian Bayes decision function defined in (4). The experimental results show the better performance of the

Table 1. Classification of training image objects

Class	1	2	3
Multispectral fusion	99.6	94.9	96.1
Probabilistic fusion	99.6	96.2	97.0

proposed approach in classes 2 (urban development) and 3 (vegetation), while maintaining an equal rate for class 1 (water).

5 Conclusion

The mathematical concept of the permanence of ratios for data fusion and the Gaussian function based Bayes classifier have been presented for object recognition in images. The proposed approach appears to be a useful tool for combining multiple probabilistic information sources. Our investigation in developing a weighted permanence of ratio based information fusion is under way.

References

1. Smith, M.I., Heather, J.P.: A review of image fusion technology in 2005. In: Proc. SPIE, vol. 5782, pp. 29–45 (2005)
2. Blum, R.S., Liu, Z. (eds.): Multi-Sensor Image Fusion and Its Applications. CRC Press, Boca Raton (2006)
3. Goshtasby, A.A., Nikolov, S.: Image fusion: Advances in the state of the art. Information Fusion 8, 114–118 (2007)
4. Pham, T.D.: An image restoration by fusion. Pattern Recognition 34, 2403–2411 (2001)
5. Pham, T.D., Tran, D.T.: Image classification by fusion for high-content cell-cycle screening. In: Gabrys, B., Howlett, R.J., Jain, L.C. (eds.) KES 2006. LNCS (LNAI), vol. 4251, pp. 524–531. Springer, Heidelberg (2006)
6. Chi, Z., Yan, H., Pham, T.: Fuzzy Algorithms: With Applications to Image Processing and Pattern Recognition. World Scientific, Singapore (1996)
7. Journel, A.G.: Combining knowledge from diverse sources: An alternative to traditional data independence hypotheses. Mathematical Geology 34, 573–595 (2002)
8. Park, N.W., Chi, K.H.: A probabilistic approach to predictive spatial data fusion for geological hazard assessment. In: Proc. IEEE Int. Geoscience and Remote Sensing Symposium, vol. 4, pp. 2425–2427 (2003)
9. Gonzalez, R.C., Woods, R.E.: Digital Image Processing. Prentice Hall, NJ (2002)

A New Wavelet–Fractal Image Compression Method

Vu Thanh Hien

Ho Chi Minh City University of Foreign Languages-Information Technology,
Ho Chi Minh, Vietnam
hienvt2000@yahoo.com

Abstract. This paper proposes a new wavelet-fractal image compression method by studying the limitation of existing wavelet based image compression methods. Initial errors occur at different levels of importance according to the frequencies of sublevel-band wavelet coefficients. Higher frequency sublevel bands would lead to larger initial errors. As a result, the sizes of sublevel blocks and super blocks would be changed according to the initial errors. The matching sizes between sublevel blocks and super blocks would be changed according to the permitted errors and compression rates.

Keywords: fractal coding, wavelet coding, image compression.

1 Introduction

The self-similarity property of images is the essence of fractal image compression methods [1] and there are great similarities among the sublevel bands with the same orientation when the wavelet transform is finished. Therefore, many researchers have utilized the similarity property of sublevel blocks for wavelet-fractal image coding [2–11].

For example, Rinaldo and Calvagno [3] have proposed an image coding approach called domain block, which is used to predict blocks in higher frequency sublevel bands by using range blocks from lower frequency sublevel bands. Davis employed a wavelet tree based fractal coding method [4–5], which combined wavelet transform and a zerotree structure proposed by Shapiro.

However, the traditional wavelet tree based fractal coding algorithm does lose sight of the property that the energy would be highly contained in low frequency sublevel bands when the wavelet transform is applied to an image and it also fails to consider the self-similarity property of fractal images. Additionally, the method still has the limitation that the speed of fractal coding is low. The new wavelet-fractal compression algorithm with a four-fork tree, proposed in this paper, presents a solution to this limitation.

2 Wavelet Tree Based Fractal Coding

The embedded zerotree coding algorithm is based on a wavelet tree [2]. With wavelet tree is meant the composition of the details related to wavelet coefficients corresponding to some image blocks in the wavelet domain \tilde{V}_j . Take the image of Lenna (256x256

pixels) for example, we can obtain ten frequency sublevel bands by three times applying wavelet transform to the images.

As shown in Fig. 1, LL_3 represents a scaled coefficient corresponding to a frequency band, whose scale is three times coarser than that of the original image. Additionally, HL_1 represent wavelet coefficients arising from separable application of vertical and horizontal filters as well as HL_2 and HL_3 . HL_1 , HL_2 and HL_3 constitute a wavelet tree. In the same way, there is another wavelet tree formed by LH_1 , LH_2 and LH_3 as well as HH_1 , HH_2 and HH_3 . And LL_3 is the root for all these wavelet trees. The components of these wavelet trees are similar, especially those of the same orientation at different scales which have the edge and texture features of self-similarity. Therefore, it would be applicable to substitute the unit of tree for the matching of domain blocks and range blocks in the spatial domain. The approaches to fractal coding could be equally utilized in the wavelet domain.

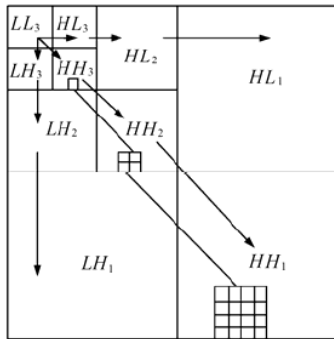


Fig. 1. Sequence distribution and paternity among the parameters of wavelet

Let $S_{K,L}^J$ be the “get-tree” operator corresponding to the “get-block” operator in fractal coding and J represents the frequency where $0 \leq J \leq N$. Let (K, L) be the relative location of the coefficient of the root with respect to the frequency in the wavelet transform domain, where $0 \leq K, L \leq 2^J$. Here $S_{K,L}^J A_w$ denotes the operator which extracts the wavelet tree, with its root at $\psi_{w,k,l}^J$, from the wavelet transform domain A_w of image A . Let $(S_{K,L}^J)^*$ be the reverse operator of $S_{K,L}^J$ which inserts the wavelet tree into the all-zero image and its root is at $\psi_{w,k,l}^J$. When a wavelet transform is applied to image A , a wavelet domain is formed. Thereby, the get-block operator $B_{K,L}^R A$ in spatial domain is replaced by the get-tree operator $S_{K,L}^{N-R} A_w$ in wavelet domain.

For example, let R be a $2^2 \times 2^2$ sublevel block in A . Since in a wavelet domain the root of wavelet tree could only exist in $N - 2$, there is only one pixel of the $2^2 \times 2^2$ sublevel block left in the wavelet domain \tilde{V}_{N-2} and the operator $B_{K,L}^R A$ is now replaced with the operator $S_{K,L}^{N-R} A_w$.

Fractal coding in the wavelet domain is similar to fractal coding in the spatial domain. It uses a wavelet supertree $S_{K,L}^{N-D} A_w$ to approach $S_{K,L}^{N-R} A_w$. The optimal approach is supposed to be

$$S_{K,L}^{N-R}A_w \approx g_{K,L} \hat{L}_P \hat{A} S_{K,L}^{N-D}A_w$$

where $g_{K,L}$ is the contraction factor. Unlike traditional fractal coding, there is no need to save the shift vector. There are two implementation methods of wavelet-fractal coding:

(i) Implementation with sublevel trees of fixed size

The wavelet domain image is partitioned into subtrees $S_{K,L}^{N-R}A_w$ of equal size, where $(K, L) \in \{(2^R m, 2^R n) | 0 \leq m, n \leq 2^{N-R}\}$. The selection of a supertree $S_{K,L}^{N-D}A_w$ is a bit different from that of the domain block in the spatial domain. If the distance between the roots of trees represents the distance between the trees, supertrees that are a pixel apart, correspond to a set of domain blocks which are 2^D in width and a domain block apart in the spatial domain.

After performing the “average and subsample” operator and “reflect and rotate” operator, the similarity is computed between the supertree and the subtree to determine whether it is the optimal supertree. Finally, the quantified wavelet coefficients and the mapping coefficients of wavelet coding are stored, including the location information of the supertree, the type of reflecting-rotating transform and the contraction factor.

When the image is decoded, the wavelet coefficients of quantified coding are recovered first and all the other wavelet coefficients are gradually recovered.

(ii) Fractal coding with four-fork partition

The four-fork partition is equally applicable to fractal coding in a wavelet domain. But it should be taken into consideration that when the root of the tree is located in the $(N-R)$ scale subtree $S_{K,L}^{N-R}A_w$, it only consists of three coefficients of three orientations. Given such condition, if no supertree is found which fits the required error in the search of matching the subtree and the supertree, the decomposition of the subtree should stop and would resume only if the frequency of the subtree had increased up to two times the original frequency. The process of decomposition would not halt until the matching error is lower than the threshold or the size of the subtree has achieved the pre-determined size.

When the image is decoded, the wavelet coefficients of hierarchically quantified coding are recovered first and then other wavelet coefficients are recovered through fractal decoding. Finally, the wavelet reverse transform is applied.

Recent researches into fractal coding based on wavelet transform have focused on the following two schemes: One is the fractal coding approach proposed by Rinaldo [3] et al. which uses blocks from low frequency sublevel bands obtained by applying wavelet transform to images in order to predict blocks in higher frequency sublevel bands. Another is the fractal predication approach based on wavelet trees, proposed by Davis [4–5]. The two algorithms work as follows:

- (i) A N -scale wavelet decomposition is applied to the image. Then the sublevel bands of the images are obtained on N scales. The image blocks in low frequency sublevel bands contain more than 90% of the energy of the original image which, thereby, should be the basis of fractal coding and should be encoded by approaches which lose little or no coding at all.

- (ii) Use LH_i , HL_i and HH_i image blocks in low frequency sublevel bands to predict image blocks in higher frequency sublevel bands of the same orientation.

When the image is decoded, images in low frequency sublevel bands are recovered first and then, given a scale from low to high, images in other frequency sublevel bands are recovered.

These wavelet-based fractal coding approaches have the following limitations, although they have taken advantage of the self-similarity among sublevel bands.

First, attention is paid to the similarity among sublevel bands, but the approaches fail to take into consideration the property that if wavelet transform is applied to images, the energy of the images would be highly contained in low frequency sublevel bands. As Shapiro [2] points out, by applying wavelet transform to images, we would obtain a large number of zero and small coefficients. Furthermore, the higher the frequency, the more zero coefficients and small coefficients are obtained. Second, the approaches fail to consider the features of self-similarity of images and the local and overall similarity among different sublevel bands. Third, there is no solution to the limitation that the speed of coding is low for fractal coding.

Given these limitations of wavelet-based fractal coding algorithms, this paper proposes a new algorithm of wavelet-fractal image compression to take effective advantage of the fine properties of wavelet coefficients corresponding to wavelet transform.

3 New Wavelet–Fractal Coding with Four–Fork Tree

The algorithm we propose is based on the proposition that, when wavelet transform is applied to images, there will be a number of zero coefficients and small coefficients. The property would be more distinct and the distribution of zero coefficients and small coefficients would be denser with regard to the higher frequency of sublevel bands. When wavelet transform is applied to image A for N times, N + 1 frequency sublevel bands would be obtained. Of all the frequency sublevel bands, except the low frequency sublevel bands from the top left corner of the image, the grey pixel scale of most areas is changed slowly into the other high frequency sublevel bands. The change of the grey pixel scale is very smooth. If we use the new wavelet-fractal coding algorithm with the four-fork tree to encode the frequency sublevel bands, this property leads to higher compression rates at relatively high speeds. Below are the two advantages of the proposed method:

- (i) By applying wavelet transform, we could obtain different frequency sublevel bands. Given the various levels of importance attached to wavelet coefficients of different frequency sublevel level bands, we assign the values of initial errors respectively – the higher the frequency sublevel band, the larger the initial error.
- (ii) The matching sizes between sublevel blocks and super blocks are dynamically changed given the initial errors in frequency of the sublevel band. In this way, if only the initial error is appointed properly, related transformation coefficients could be obtained by utilizing the similarity among many large-size sublevel blocks and super blocks. If these large sized sublevel blocks cannot satisfy these requirements, we can change the length of the edge within the range of $2^J, 2^{J-1}, \dots$. The corresponding superblock is equally constructed. The matching sizes between

sublevel blocks and super blocks are dynamically changed according to initial error of the frequency of the sublevel band and the compression rate.

This algorithm works as follows:

- (i) DWT is applied to image A for J times. Through wavelet transform, the image A is partitioned into J+1 sublevel bands whose edges are in the range of 2^{N-1} , 2^{N-2} , ..., 2^{N-J} . We use the approach which loses little or no coding to encode LL_0 from the low frequency sublevel bands, which contains more than 90% of the energy of the original image. To obtain a higher compression rate, we could use the basic fractal compression algorithm of Jacquin. In order to reduce the compression time, we could classify LL_0 according to the entropy and deviation of the grey scale.
- (ii) LH_0 , HL_0 and HH_0 images in frequency sublevel band, have the same size of $2^{N-J} \times 2^{N-J}$. Let ϵ_j be an acceptable error of the matching sizes between the super block and the sublevel block in the fractal compression coding process. $2^{R_{max}^j}$ denotes the maximum length of the edge while $2^{R_{min}^j}$ is the minimum length of the edge where $2^{R_{max}^j} \leq 2^{N-J-1}$, $2^{R_{min}^j} \geq 2^2$.
- (iii) To begin with, the sublevel bands LH_0 , HL_0 and HH_0 are partitioned into disjointed $2^{R_{max}^j} \times 2^{R_{max}^j}$ subblocks R_i^j . Let $B_{K,L}^{R_{max}^j}(LH_0)$ be the get-block operator. (K,L) represents the location on the top left corner. Take the sublevel band LH_0 for example. We search for its corresponding super block located in $\pi(K,L)$ by using the “average and subsample” operator and the “reflect and rotate” operator. Then we compute the minimum error between $L_{P(K,L)}(LH_0)$ and $B_{K,L}^{R_{max}^j}(LH_0)$, that is d_{min}^p . If d_{min}^p is smaller than ϵ_j , we save the location information of both the sublevel block and the super block as well as the type of reflecting-rotating-transform and the contraction factor into a compression document. In this way, the coding of R_i^j is completed. Otherwise, if d_{min}^p is larger than ϵ_j , R_i^j will be partitioned into four sublevel blocks and so will the super block. The search for the most similar sublevel block and super block would be on until the self-similarity transform is smaller than ϵ_j for sublevel blocks of all sizes and their corresponding super blocks.
- (iv) For the other sublevel bands such as I_1, I_2, \dots, I_j , where $I_i = \{LH_i, HL_i, HH_i\}$ $1 \leq i \leq J$, the wavelet coefficients become less and less significant with a scale increase. So the value of the initial error could become increasingly larger such as $2\epsilon_j, 4\epsilon_j, \dots, 2^{j-1}\epsilon_j$. $2^{R_{max}^j}$ denotes the maximum edge length while $2^{R_{min}^j}$ is the minimum length of the edge where $2^{R_{max}^j} \leq 2^{N-J-1}$, $2^{R_{min}^j} \geq 2^2$. Then we could repeat steps 2 and 3 to complete the coding. In this way, the image is partitioned into small blocks when the change of image is complex while the blocks would be large when the change is smooth. Because the coefficients of frequency of the images at the sublevel bands are small and their change is smooth, the blocks are partitioned into large ones. As a result, the compression time is short.

When the image is decoded, iteration is applied first to recover the wavelet coefficients. Then by applying a wavelet transform, we recover the original image.

4 Experimental Results

Table 1 lists the experimental results of encoding the image of Lenna by the algorithm proposed in this paper using Spline wavelet bases which are compared with the results of fractal coding based on a wavelet tree. As shown in Table 1, although the PSNR is reduced, the distortion rate is not high and the compression rate and the speed of coding have clearly increased. It demonstrates that a better coding result can be obtained.

Table 1. Comparison of compression methods

Figure number	Method	Compression rate	PSNR (dB)	Time (s)
2a	Fractal coding based on wavelet tree	17.3	31.3	148
2b	New wavelet-fractal coding of four-fork tree	25.6	29.3	36

Table 2 lists the experimental results of different images.

Table 2. The experimental results of different images

Method	Image	Compression rate	PSNR (dB)	Time (s)
Fractal coding based on wavelet tree	woman	19.3	31.5	151
	peppers	16.9	30.4	142
	house	18.5	30.9	145
New wavelet-fractal coding of four-fork tree	woman	26.3	30.1	32
	peppers	22.6	30.3	37
	house	24.7	29.8	38

Table 3 shows some comparative results for our method, the pure Fisher’s accelerated fractal coding (QPIFS) and the pure SPIHT wavelet coding.

Table 3. The comparison of our method and other methods

Method		Lena	woman	peppers	house
QPIFS	PSNR (dB)	28.8	29.6	30.8	28.9
	Time (s)	76	73	68	71
	Compression rate	25	26.3	24	25
SPIHT	PSNR (dB)	30.4	29.8	29.9	29.7
	Time (s)	40	31	45	42
	Compression rate	25.2	24.5	26	23.8
our method	PSNR (dB)	29.3	30.1	30.3	29.8
	Time (s)	36	32	37	38
	Compression rate	25.6	26.3	22.6	24.7



Fig. 2. (a) Fractal coding based on wavelet tree (b) New wavelet-fractal coding of four fork tree



Fig. 3. Test images. (a) woman (b) peppers (c) house.

5 Conclusion

Although wavelet tree based fractal coding algorithms have taken advantage of self-similarity among sublevel bands, they fail to take into consideration the property that when a wavelet transform is applied to an image, its energy would be largely contained in low frequency sublevel bands. Equally, they fail to consider features of fractal image-self-similarity. As well, these algorithms still have the limitation that the speed of fractal coding is low. The new wavelet-fractal compression algorithm with a four-fork tree, proposed in this paper, has a solution to this limitation. Theoretical analysis and experimental results demonstrate that, compared with the classical wavelet tree based algorithms of fractal image compression, this algorithm clearly increases the compression rate and the speed of encoding without reducing PSNR and the quality of decoded images.

References

1. Jacquin, A.E.: Fractal image coding: a review. *IEEE Transaction on Image Processing* 81(10), 1451–1465 (1993)
2. Shapiro, J.: Embedded image coding using zerotrees of wavelet coefficients. *IEEE Trans. Signal Processing* 41(12), 3445–3462 (1993)
3. Roberto, R., Giancarlo, C.: Image coding by block prediction of multiresolution subimages. *IEEE Transactions on Image Processing* 4(7), 909–920 (1995)

4. Davis, G.M.: Self-quantization of wavelet subtrees: a wavelet based theory of fractal image compression. In: Proc. Data Compression Conf. USA, pp. 232–241 (1995)
5. Davis, G.M.: A wavelet-based analysis of fractal image compression. *IEEE Trans. on Image Processing* 7(2), 141–154 (1998)
6. Kim, S.H., Jang, I.H.: Image coding using wavelet-based fractal approximation. *IEICE Transactions on Information and Systems* 85(10), 1723–1726 (2002)
7. Ghazel, M., Freeman, G.: Fractal-wavelet image denoising. In: IEEE International Conference Conference on Image Processing, vol. 1, pp. 836–839 (2002)
8. Xie, X., Ma, Z.M.: Fractal predictive image coding based on zerotrees of wavelet coefficients. *Journal of Image and Graphics* 5A(11), 920–924 (2000)
9. Xie, Y.H., Fu, D.S.: A fractal image coding algorithm research based on wavelet transformation. *Journal of Image and Graphics* 8A(7), 839–842 (2003)
10. Wang, D.F., Jiang, W.: Fractal image coding combined with wavelet subtree. *Systems Engineering and Electronics* 27(6), 1120–1122 (2005)
11. Zhao, J., Pan, J.S., Chen, G.H.: Study on application of combination of wavelet with fractal in image processing. *Computer Engineering* 31(1), 29–52 (2005)
12. Zhou, Y.-M., Zhang, C., Zhang, Z.-K.: Fast hybrid fractal image compression using an image feature and neural network. *Chaos, Solitons & Fractals* 37, 623–631 (2008)
13. Kunze, H.E., La Torre, D., Vrscay, E.R.: Random fixed point equations and inverse problems using collage method for contraction mappings. *Journal of Mathematical Analysis and Applications*, 1116–1129 (2007)
14. Christophe, E., Mailhes, C., Duhamel, P.: Hyperspectral Image Compression: Adapting SPIHT and EZW to Anisotropic 3-D Wavelet Coding. *IEEE Transactions on Image Processing* 17(12) (December 2008)
15. Chang, C.-L., Girod, B.: Direction-Adaptive Discrete Wavelet Transform for Image Compression. *IEEE Transactions on Image Processing* 16(5) (May 2007)

Urban Vehicle Tracking Using a Combined 3D Model Detector and Classifier

Norbert Buch, Fei Yin, James Orwell, Dimitrios Makris, and Sergio A. Velastin

Digital Imaging Research Centre, Kingston University, Penrhyn Road,
Kingston upon Thames, KT1 2EE, United Kingdom
{norbert.buch, fei.yin, j.orwell, d.makris,
sergio.velastin}@kingston.ac.uk

Abstract. This paper presents a tracking system for vehicles in urban traffic scenes. The task of automatic video analysis for existing CCTV infrastructure is of increasing interest due to benefits of behaviour analysis for traffic control. Based on 3D wire frame models, we use a combined detector and classifier to locate ground plane positions of vehicles. The proposed system uses a Kalman filter with variable sample time to track vehicles on the ground plane. The classification results are used in the data association of the tracker to improve consistency and for noise suppression. Quantitative and qualitative evaluation is provided using videos of the public benchmarking i-LIDS data set provided by the UK Home Office. Correctly detected tracks of 94% outperform a baseline motion tracker tested under the same conditions.

Keywords: vehicle tracking, visual surveillance, motion estimation, 3D models, vehicle classification, urban traffic, performance evaluation.

1 Introduction

In recent years, there has been an increased scope for automatic analysis of urban traffic activity. This is due in part to the additional numbers of cameras and other sensors, the enhanced infrastructure and consequent accessibility and also the advancement of analytical techniques to process the video data. Monitoring objectives include the detection of traffic violations (illegal turns, one way streets, *etc.*) and the gathering of statistics about the type of road users. Using general purpose surveillance cameras, the classification of vehicles is a demanding challenge (see [9, 8, 12, 4]). Compared to most examples in image retrieval problem, the quality of surveillance data is generally poor and the range of operational conditions (night-time, inclement and changeable weather that affects the auto-iris) require robust techniques which need to be immune to errors in obtaining road users' silhouettes. Those silhouettes extracted by foreground analysis are the input to our classifier. The classification process is based on 3D models for vehicles to give robustness against foreground noise and can be restricted to an active region of the camera view (*e.g.* lanes). This allows human operators to configure monitoring objectives. The classified vehicles are tracked on the ground plane over time using a Kalman filter for variable time steps. Tracking performance is evaluated using the

framework of Yin *et al.* [13] and compared to a state of the art OpenCV blob tracker [11] operating on the same video data.

Our novel contributions are firstly the extension of our 3D vehicle detector and classifier by tracking on the ground plane. We derive a variable sample rate Kalman filter to accommodate missed observations. The classification of vehicles is used during tracking due to our novel approach of classifying before tracking. Secondly, our tracking evaluation framework [13] is used to generate rich performance figures based on ground truth containing image bounding boxes. Thirdly, the performance of the 3D model based ground plane tracker is compared to a state of the art blob tracker.

The remainder of the paper is organised as follows: Section 2 introduces the detector and classifier used. The application of Kalman filtering to the classification results is demonstrated in section 3. Introduction to the evaluation framework and results are given in section 4. Section 5 concludes the paper.

1.1 Related Work

This review firstly introduces detection and tracking systems and continues with performance evaluation frameworks. Vehicle tracking in urban environments is performed in [12]. However, only a single 3D model for cars is used to estimate a vehicle constellation per frame with optimisation solved with a Markov Chain Monte Carlo (MCMC) algorithm. The reported detection rates are 96.8% and 88% for two videos, which are limited to single size vehicles. The paper of Morris and Trivedi [9] presents a combined tracking and classification approach for side views of highways which is an extension to [8]. A single Gaussian background model is used for foreground segmentation. Classification and tracking accuracy was increased by combining tracking and classification. A Kalman filter is used to track the foreground regions based on the centroids in the image plane only. The OpenCV blob tracker [11] used as baseline here works in a similar fashion. The field of generic object recognition recently expanded towards surveillance applications. Good examples are Leibe *et al.* [6,7] for vehicle and pedestrian detection. Performance however, is not yet comparable to state of the art surveillance systems for this specific task.

Performance evaluation has played an important role in developing, assessing and comparing object tracking algorithms. Lazarevic-McManus *et al.* [5] evaluated performance of motion detection based on ROC-like curves and the F-measure. The latter allows comparison using a single value domain, but is mainly designed to operate on motion detection rather than tracking. There is a significant body of work dealing with evaluation of both motion detection and tracking. Needham and Boyle [10] proposed a set of metrics and statistics for comparing trajectories to account for detection lag, or constant spatial shift. However, taking only the trajectory (a set of points over time) as the input of evaluation may not give sufficient information about how precise the tracks are, since the size of the object is not considered. Bashir and Porikli [1] use the spatial overlap of ground truth and system bounding boxes which is unbiased towards large objects. However they are counted per frame, which is justified when the objective is object detection. In object tracking, counting true positive (TP), false positive (FP) and false negative (FN) tracks is a more natural choice which is consistent with the expectations of surveillance end-users. Brown *et al.* [3] suggests a framework for matching of system track centroids and an enlarged ground truth bounding box which favours tracks of large objects.

2 Detection and Classification Using 3D Models

Joint detection and classification is performed using 3D wire frame models for vehicles with calibrated cameras. As indicated in the block diagram in Figure 1, the detector uses a Gaussian Mixture Model (GMM) for motion estimation with subsequent closed contour retrieval to generate motion silhouettes for an input video frame. Those motion silhouettes are used to generate vehicle hypotheses. The classifier matches 3D wire frame models (see) with the motion silhouettes. To validate the hypotheses, the normalised overlap area of motion silhouettes and projected model silhouettes is calculated. Full details on the classifier can be found in a previous paper [4]. The output of the classifier are class labelled ground plane positions of vehicles. On frame to frame detection and classification of four classes, the classifier precision is 96.1% with a total system recall of 90.4% at a precision of 87.9%. Section 5 gives tracking evaluation results on the same video set.

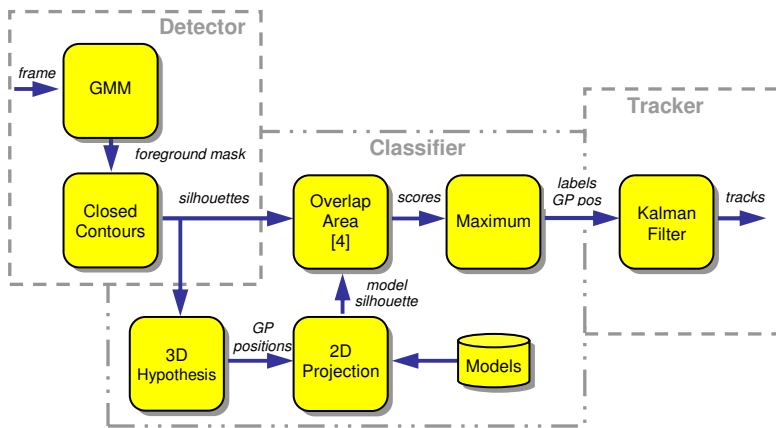


Fig. 1. Block diagram of detector with 3D classifier and subsequent tracker

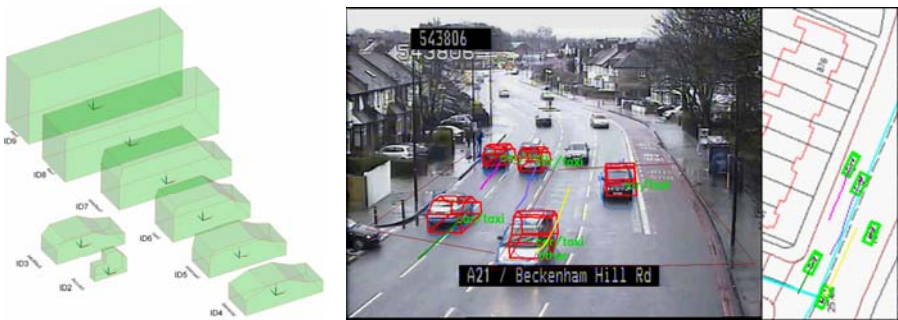


Fig. 2. Left: 3D wire frame models used for the classifier. Right: Example of detection and classification with ground plane tracking. The wire frame projection in red is used to estimate the bounding box for tracked vehicles.

3 Tracking

Tracking introduces temporal consistency to the detection and classification result of the previous section. Our novel contribution is the extension of the classifier by a Kalman filter with variable sample rate. The detector with joint classifier may reject valid vehicles in some frames due to noise, which requires the Kalman filter to operate on variable time intervals. Tracking is performed on the ground plane of the scene, which simplifies behaviour analysis like bus lane monitoring. We use the standard formulation of the Kalman filter for a constant velocity model of vehicles

$$\mathbf{x}_k = \mathbf{F}\mathbf{x}_{k-1} + \mathbf{B}\mathbf{u}_k + \mathbf{w}_k \quad \mathbf{z}_k = \mathbf{H}\mathbf{x}_k + \mathbf{v}_k \quad \text{with } \mathbf{u}_k = \mathbf{0} \quad (1)$$

with state vector $\mathbf{x}_k = [v_x, x, v_y, y]^T$ and the measurement vector $\mathbf{z}_k = [x, y]^T$. All time and speed related constants for the filter are based on seconds rather than the sample rate or frame rate. The ground plane coordinates are in metres, all noise and position estimates are in metres or meters per second. The above is valid, if the integration constant T_0 from speed to position in the transition matrix \mathbf{F} is defined in seconds

$$\mathbf{F} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ T_0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & T_0 & 1 \end{bmatrix}. \quad (2)$$

The only conditions to operate the Kalman filter at variable sample rate is to update T_0 in the transition matrix \mathbf{F} constantly. For prediction steps, T_0 is the time between the last update step of the filter and the current time. The state prediction $\hat{\mathbf{x}}_{k|k-1}$ and the error covariance prediction $\mathbf{P}_{k|k-1}$ is therefore estimated for the correct time. If a measurement is available, the update step is performed with the same transition matrix \mathbf{F} . If no measurement is available, not update is performed. Future prediction steps will be performed with increasing time T_0 until an update takes place. Tracks can be discarded if the predicted error covariance $\mathbf{P}_{k|k-1}$ grows beyond a threshold.

The parameters for the filter are as follows. The process noise \mathbf{w} is set to $1.1m/s$ for velocity and $0.7m$ for position. Those values can be derived from the expected acceleration of vehicles. The measurement noise is $\mathbf{v} = 2m$ corresponding to the detection grid. The initial error covariance \mathbf{P} is set to $3m/s$ for velocity and $1m$ for position. The initial position state corresponds to the detection position with zero velocity. The velocity is updated during the second detection using the first motion vector. Observations $\mathbf{m}_{i,k}$ are associated with tracks based on the distance d_{ij} between the observation $\mathbf{m}_{i,k}$ and the prediction $\hat{\mathbf{x}}_{j,k|k-1}$ normalised by the diagonal elements of the predicted error covariance $\mathbf{P}_{k|k-1}$. Changes in the model- id between the last observation of a track id_i and the current observation id_j are penalised. The total number of model- ids is 10. This novel approach is possible because our system performs classification before the tracking.

$$d_{ij} = \sqrt{\left[(x_i - x_j) P_x^{-1} \right]^2 + \left[(y_i - y_j) P_y^{-1} \right]^2 + \frac{1}{10} |id_i - id_j|} \quad (3)$$

4 Evaluation

The object tracking performance is demonstrated by comparing our tracker with a baseline tracker (OpenCV blob tracker [11]). The OpenCV tracker uses an adaptive mixture of Gaussians for background estimation, connected component analysis for data association and Kalman filtering for tracking blob position and size. We use the i-LIDS bench- marking video data set provided by the UK Home Office [2] for evaluation. We run the tracker on the following sequences of the parked car data set scene 1 (PVTRA10xxxx): 1a03, 1a07, 1a13, 1a19, 1a20, 2a05, 2a10 and 2a11. Those videos contain overcast, sunny, changing weather conditions and camera saturation.

We propose a rich set of metrics such as Correct Detected Tracks, False Detected Tracks and Track Detection Failure to provide a general overview of the system's performance. Track Fragmentation shows whether the temporal and spatial coherence of tracks is established. ID Change is useful to test the data association module of the system. Latency indicates how quick the system can respond to an object entering the camera view, and Track Completeness how complete the object has been tracked. Metrics such as Track Distance Error and Closeness of Tracks indicate the accuracy of estimating the position, the spatial and the temporal extent of the objects respectively. More details about this evaluation framework can be found in Yin *et al.* [13].

4.1 Qualitative Results

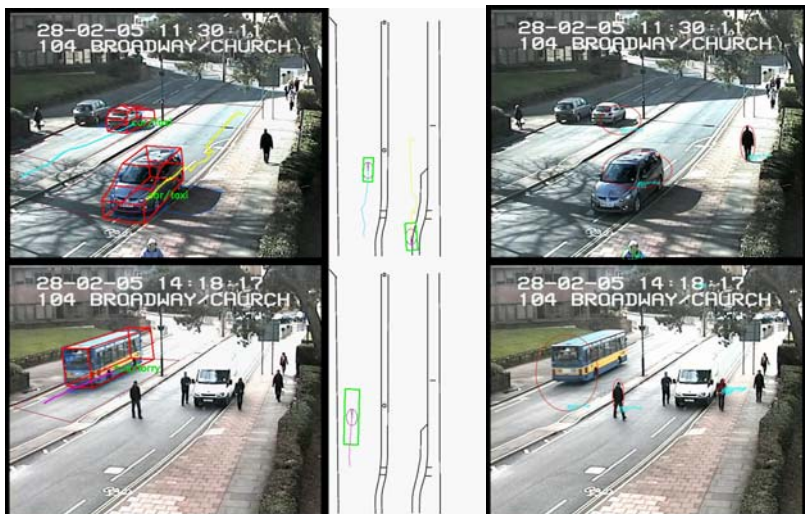


Fig. 3. Correct detected tracks inside the active regions of interest (dark red boxes). Left: the proposed system with corresponding ground plane tracks. Right: OpenCV tracker result.



Fig. 4. The second car is missed due to occlusion between the vehicles. The proposed classifier on the left correctly locates the first car. The OpenCV tracker merged both cars with a large bounding box at a central position.

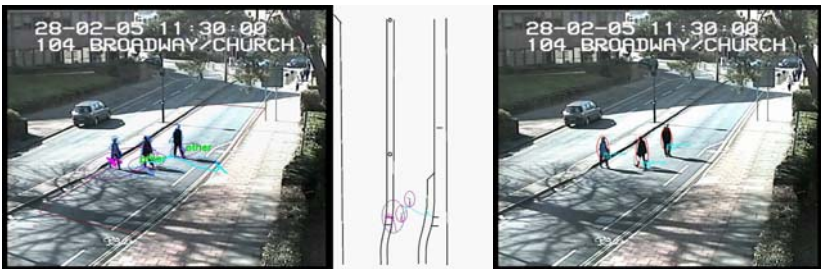


Fig. 5. Pedestrians are correctly rejected as “other” class by the proposed classifier and detected by the OpenCV tracker

4.2 Quantitative Results

The ground truth used for evaluation is provided with the i-LIDS data set. It is of limited duration within the videos and does not include pedestrians on the road. The evaluation was constrained to the two regions of interest on the road (dark red boxes in Figure 3) for both trackers. The full results are provided in Table 1 indicating that the proposed system outperforms the OpenCV tracker on high level metrics such as correct detected tracks, track detection failure, false detected tracks and track fragmentation. This can mainly be attributed to the additional prior information from using 3D models to classify the content of the input video.

For metrics that evaluate the motion segmentation such as track closeness and distance error, both trackers have similar performance, which can be explained by the similar background estimation method. The track closeness of the proposed system is better than the baseline due to 3D models which are more robust against shadows, which can be observed for the bus in Figure 3 and the occluded car in Figure 4. The extent of the projected wire frame model is used as bounding box for the proposed system. The false detected tracks of the OpenCV tracker are high due to systematic detection of pedestrians, which can not be classified. Refer to Figure 5 for an example. The proposed system detected 94% of the ground truth tracks compared to 88% of the base line. Our system has half of the track detection failures compared to the

Table 1. Tracking results

<i>Metrics</i>	<i>proposed Tracker</i>	<i>OpenCV blob Tr.</i>
Number of Ground truth tracks	100	100
Number of system tracks	144	203
Correct detected tracks	94	88
Track detection failure	6	12
False detected tracks	27	90
Latency [frames]	5	5
Track fragmentation	8	18
Average track Completeness [time]	64%	55%
ID change	10	3
Average track closeness [bbox overlap]	54%	35%
Standard Deviation of closeness	20%	13%
Average distance error [pixels]	22	21
Standard Deviation of distance error	19	15

base line. The higher detection rate can be explained by a more sensitive background estimation producing more complete and additional noise detections. However, the classification stage rejects many ambiguous detections. Id change can occurs if a track of an object leaving is continued for a new object. This is worse for the proposed system compared to the OpenCV tracker, because the tracker is more persistent, occasionally wrongly continuing a track but therefore generating much less track fragmentations.

5 Conclusions and Future Work

We proposed a novel system for detection, classification and ground plane tracking of vehicles in surveillance videos. The proposed system is evaluated on the i-LIDS data set against the state of the art OpenCV blob tracker. Our system performs similar for motion related metric but outperforms the baseline for high level metric like detected tracks 94% and missed tracks 6. This indicates superior performance in the camera view with the additional benefit of gaining group plane locations. This can be essential to solve surveillance tasks like enforcing bus lane restrictions.

Future work can be the evaluation of the classes of tracks and the ground plane positions. Both require a significant amount of ground truth. Regarding the detector and classifier, avoiding the reliance on motion estimation would be beneficial for more robustness against lighting changes and camera saturation. There is the opportunity to post process completed tracks for retrospective behaviour analysis.

Acknowledgements

We are grateful to the Directorate of Traffic Operations at Transport for London for funding the work on classification and tracking and to BARCO View, Belgium for funding the work on tracking evaluation.

References

- [1] Bashir, F., Porikli, F.: Performance evaluation of object detection and tracking systems. In: IEEE Int. W. on Performance Evaluation of Tracking and Surveillance, PETS 2006 (2006)
- [2] Home Office Scientific Development Branch. Imagery library for intelligent detection systems i-lids,
<http://scienceandresearch.homeoffice.gov.uk/hosdb/cctv-imaging-technology/video-based-detection-systems/i-lids/> (accessed December 19, 2008)
- [3] Brown, L.M., Senior, A.W., Tian, Y.L., Connell, J., Hampapur, A., Shu, C.-F., Merkl, H., Lu, M.: Performance evaluation of surveillance systems under varying conditions. In: IEEE Int'l Workshop on Performance Evaluation of Tracking and Surveillance, Colorado, January 2005, pp. 1–8 (2005)
- [4] Buch, N., Orwell, J., Velastin, S.A.: Detection and classification of vehicles for urban traffic scenes. In: International Conference on Visual Information Engineering, VIE 2008, July 2008, pp. 182–187. IET (2008)
- [5] Lazarevic-McManus, N., Renno, J.R., Makris, D., Jones, G.A.: An object-based comparative methodology for motion detection based on the f-measure. *Computer Vision and Image Understanding*. Sp. Is. on Intelligent Visual Surveillance, 74–85 (2007)
- [6] Leibe, B., Cornelis, N., Cornelis, K., Van Gool, L.: Dynamic 3d scene analysis from a moving vehicle. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2007, June 2007, pp. 1–8 (2007)
- [7] Leibe, B., Schindler, K., Cornelis, N., Van Gool, L.: Coupled object detection and tracking from static cameras and moving vehicles. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30(10), 1683–1698 (2008)
- [8] Morris, B., Trivedi, M.: Robust classification and tracking of vehicles in traffic video streams. In: Intelligent Transportation Systems Conference, ITSC 2006, pp. 1078–1083. IEEE, Los Alamitos (2006)
- [9] Morris, B., Trivedi, M.: Improved vehicle classification in long traffic video by cooperating tracker and classifier modules. In: AVSS 2006: Proceedings of the IEEE International Conference on Video and Signal Based Surveillance, USA, p. 9 (2006)
- [10] Needham, C.J., Boyle, R.D.: Performance evaluation metrics and statistics for positional tracker evaluation. In: Crowley, J.L., Piater, J.H., Vincze, M., Paletta, L. (eds.) ICVS 2003. LNCS, vol. 2626, pp. 278–289. Springer, Heidelberg (2003)
- [11] OpenCV. Open source computer vision library,
<http://sourceforge.net/projects/opencvlibrary> (accessed December 19, 2008)
- [12] Song, X., Nevatia, R.: Detection and tracking of moving vehicles in crowded scenes. In: IEEE W. on Motion and Video Computing, WMVC 2007, p. 4 (2007)
- [13] Yin, F., Makris, D., Velastin, S.A.: Performance evaluation of object tracking algorithms. In: 10th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, PETS 2007, Rio de Janeiro (October 2007)

UBIAS – Type Cognitive Systems for Medical Pattern Interpretation

Lidia Ogiela¹, Marek R. Ogiela², and Ryszard Tadeusiewicz²

AGH University of Science and Technology

¹Faculty of Management,

²Institute of Automatics

Al. Mickiewicza 30, PL-30-059 Krakow, Poland

{logiela,mogiela,rtad}@agh.edu.pl

Abstract. This paper presents some important aspects of cognitive informatics operated by cognitive processes in the human mind and implemented to new generation IT systems. This paper presents especially a selected class of cognitive categorization systems called UBIAS (*Understanding Based Image Analysis Systems*). The UBIAS systems are especially dedicated to support analysis of data recorded in the form of images for example medical images. Cognitive categorization systems operate by executing a particular type of human thought, cognitive and analysis processes which take place in the human mind and which ultimately lead to making an in-depth description of the analysis and interpreting reasoning process. The most important element in this analysis and reasoning process is that it occurs both in the human cognitive process and in the system's information process that conducts the in-depth interpretation and analysis of data.

Keywords: Cognitive informatics, reasoning processes and analysis of medical images, information systems, UBIAS systems, pattern classification.

1 Introduction

Analysis, interpretation and reasoning processes were used to build and describe new classes of cognitive categorisation systems which are to execute in-depth analyses and reasoning processes on the data being interpreted. The main subject of this publication is to present a selected class of cognitive categorisation systems – UBIAS (Understanding Based Image Analysis Systems) – which support analyses of data recorded in the form of images [7-9]. Cognitive categorisation systems operate by executing a particular type of thought, cognitive and reasoning processes which take place in the human mind and which ultimately lead to making an in-depth description of the analysis and reasoning process.

The most important element in this analysis and reasoning process is that it occurs both in the human cognitive/thinking process and in the system's information/reasoning process that conducts the in-depth interpretation and analysis of data [9]. It should be added that this process is based on cognitive resonance (Fig. 1) which occurs during the examination process, and which forms the starting point for the

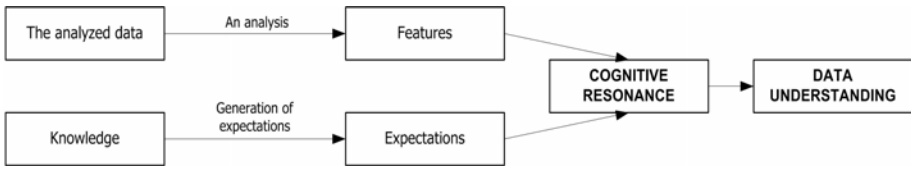


Fig. 1. Cognitive resonance in data understanding process

process of data understanding consisting in extracting the semantic information and the meaning contained in the analysed type of data that makes reasoning possible.

Cognitive resonance is an attempt to compare and distinguish certain similarities and differences between the set of analysed data and the set represented by a knowledge base. The set of data containing the analysed data group is subjected to a process of broadly-understood analysis, which means analysing the form, content, meaning, shape and the like. This analysis makes it possible to extract certain significant features of the analysed data. At the same time, the set of knowledge collected (possessed) by the system is used to generate certain expectations as to the substantive content of the analysed data. These expectations are then compared to the features of the analysed data extracted during the analysis process. When the features and expectations are compared, cognitive resonance occurs; its essence is that it indicates the similarities that appear between the analysed dataset and the generated set of expectations about the possible results of the knowledge acquired by the system. These similarities are revealed during the comparative analysis conducted by the system, in the course of which the analysed data is subjected to the phenomenon of understanding. The reasoning process which forms the result of the understanding process is an indispensable factor for the correct data analysis. If it did not occur, it would become impossible to forecast and reason as to the future of the phenomenon being studied. So conducting the analysis without the reasoning process could actually lead to impoverishing the entire analysis process, as it would be limited only to understanding the reasons why the analysed phenomenon occurred, but without a chance of determining its further development.

2 Cognitive Informatics in UBIAS Systems

Cognitive informatics in selected type cognitive systems using analysis and interpretations processes operated by human mind processes. Especially UBIAS systems which describe more and more human mind processes and their application in automatics processes and computer data analysis. In this paper we describe example of UBIAS systems dedicated to medical image cognitive analysis. We analysed images of food bone. All the analysed images of foot bones were, before their proper recognition, subject to segmentation and filtration procedures [9]. Structures shown in this way were then subject to cognitive analysis stages using the grammar described below.

So example of using the cognitive interpretation of image-type data to analyse data depicting foot bone pathologies is an analysis of images acquired in the dorsoplantar projection.

For such a projection, the appropriate set of foot bone names:

- heel (c)
- os naviculare (on)
- os cuboideum (oc)
- os cuneiforme mediale (ocm)
- os cuneiforme intermedium (oci)
- os cuneiforme laterale (ocl).

The food bone names was adopted and a linguistic description was introduced which presents the foot bone skeleton corresponding to the correct anatomy of this part of the lower extremity. In order to analyse X-rays of foot bones in the dorsoplantar projection, it was necessary to define a graph representation of these bones showing numbers consistent with the neighbourhood relationships of these structures. Such a definition is formulated based on a description of a graph of special topological relationships between particular elements of the graph. A graph of spatial relationships is show in Fig. 2. with a graph of special relationships to build a graph containing the numbers of adjacent foot bones for an dorsoplantar projection.

In order to analyse disease lesions of the foot bone, the following grammar has been proposed:

$$G_{dp} = (N, \Sigma, \Gamma, ST, P)$$

where:

The set of non-terminal labels of apexes:

$N = \{ST, \text{CALCANEUS}, \text{OS NAVICULARE}, \text{OS CUBOIDEUM}, \text{OS CUNEIFORME MEDIALE}, \text{OS CUNEIFORME INTERMEDIUM}, \text{OS CUNEIFORME LATERALE}, M1, M2, M3, M4, M5\}$

The set of terminal labels of apexes:

$\Sigma = \{s, t, u, v, w, x, y, c, on, oc, ocm, oci, ocl, m1, m2, m3, m4, m5\}$

Γ – the graph shown in Fig.2

ST - The start symbol

P – a finite set of productions shown in Fig. 2.

Defining such elements of the grammar is aimed at specifying a set of grammatical rules allowing all cases of images showing the correct structure of foot bones to be interpreted. It should be kept in mind that this is a different set of grammatical rules for every projection. Figure 3 shows a set of graphs defining the correct structure of foot bones visible in the dorsoplantar projection.

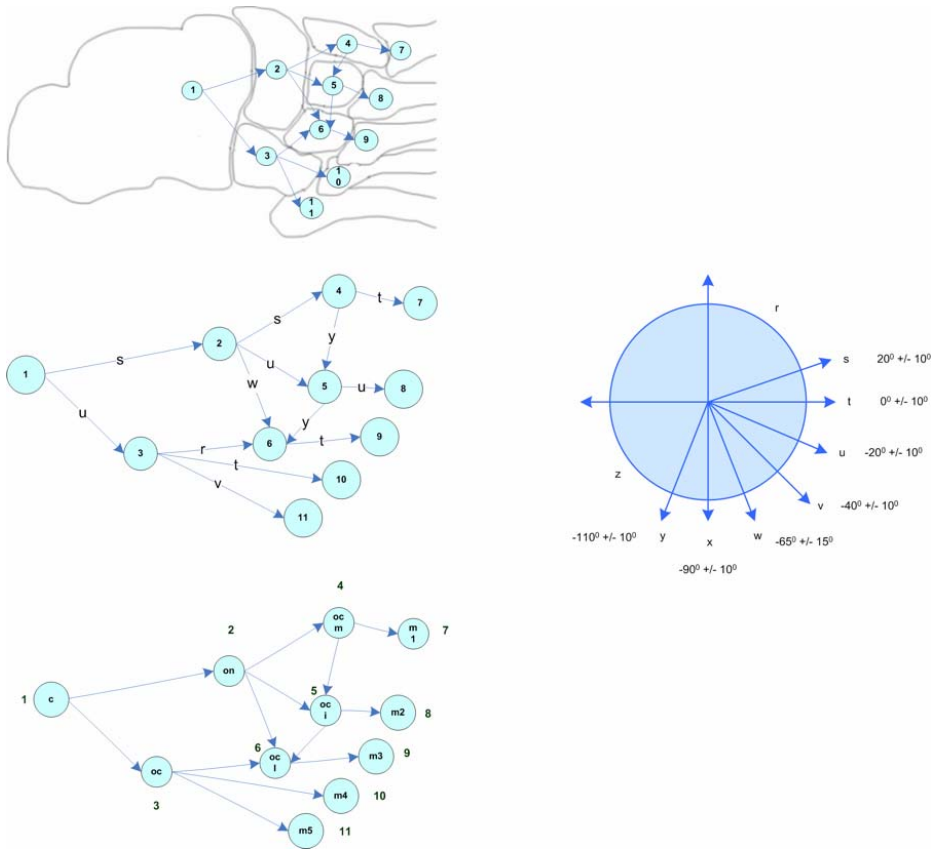


Fig. 2. The definition of an EDG graph describing the foot bone skeleton in the dorsoplanar projection. A graph of spatial relationships for the dorsoplanar projection for foot imaging. A graph with numbers of adjacent bones marked based on the graph of spatial relationships. Interrelations between particular elements of the structure of foot bones for the dorsoplanar projection.

Determining the correct relationships and the correct structure of foot bones enables UBIAS systems to conduct a meaning analysis. For such projections of foot images, these analyses can generate results of reasoning about and interpreting selected types of fractures and pathological situations whose examples are shown in Fig. 3.

The presented type of automatic understanding of image-type data for interpreting and analysing X-rays of foot bones in the dorsoplanar projection was aimed at detecting various types of fractures and irregularities appearing in the structure of the bone and at detecting lesions – haematomas.

Additional the cognitive interpretation of image-type data to analyse data depicting foot bone pathologies is an analysis of images acquired in the proper dorsoplanar projection – all foot bone image in dorsoplanar projection.

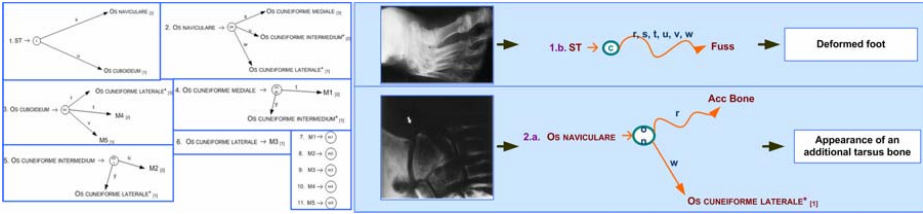


Fig. 3. A set of grammar rules showing the healthy structure of foot bones including their numbers for the dorsoplantar projection, and the automatic understanding of foot bone lesions detected by the UBIAS system in the dorsoplantar projection

The new graph representation showing numbers consistent with the neighbourhood relationships of these structures and a graph of special topological relationships between particular elements of the graph was defined. A graph of spatial relationships is shown in Fig. 4. with a graph of spatial relationships to build a graph containing the numbers of adjacent foot bones for an dorsoplantar projection.

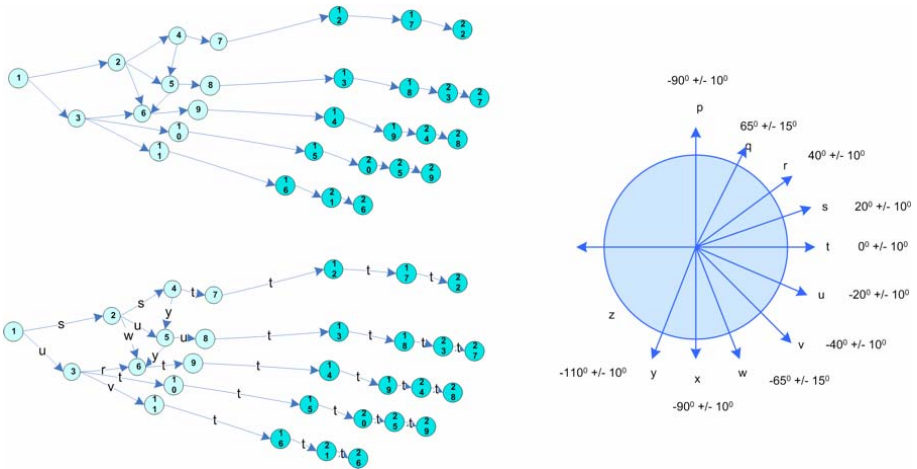


Fig. 4. The definition of an EDG graph describing the foot bone skeleton in the proper dorsoplantar projection, a graph of spatial relationships, a graph with numbers of adjacent bones marked based on the graph of spatial relationships

In order to analyse disease lesions of the foot bone, the following grammar has been proposed:

$$G_{dp2} = (N, \Sigma, \Gamma, ST, P)$$

where:

The set of non-terminal labels of apexes:

$N = \{ST, TALUS, CUBOIDEUM, NAVICULARE, LATERALE, MEDIALE, INTERMEDIUM, SES1, SES2, TM1, TM2, TM3, TM4, TM5, MP1, MP2, MP3, MP4,$

MP5, PIP1, PIP2, PIP3, PIP4, PIP5, DIP2, DIP3, DIP4, DIP5, TPH1, TPH2, TPH3, TPH4, TPH5, ADD1, ADD2, ADD3, ADD4, ADD5, ADD6, ADD7, ADD8, ADD9, ADD10, ADD11, ADD12, ADD13, ADD14}

The set of terminal labels of apexes:

$\Sigma = \{c, t, cu, n, cl, cm, ci, s1, s2, tm1, tm2, tm3, tm4, tm5, mp1, mp2, mp3, mp4, mp5, pip1, pip2, pip3, pip4, pip5, dip2, dip3, dip4, dip5, tph1, tph2, tph3, tph4, tph5, add1, add2, add3, add4, add5, add6, add7, add8, add9, add10, add11, add12, add13, add14\}$

$\Gamma = \{p, q, r, s, t, u, v, w, x, y, z\}$ – the graph shown in Fig.4

ST–The start symbol

P–set of productions.

Defining such proper elements of the grammar showing the correct structure of foot bones in dorsoplantar projection. It should be kept in mind that this is a different set of grammatical rules for every projection. Figure 5 shows a set of graphs defining the correct structure of foot bones visible in the dorsoplantar projection. Determining the correct relationships and the correct structure of foot bones enables UBIAS systems to conduct a meaning analysis. For such projections of foot images, these analyses can generate results of reasoning about and interpreting selected types of fractures and pathological situations whose examples are shown in Fig. 5

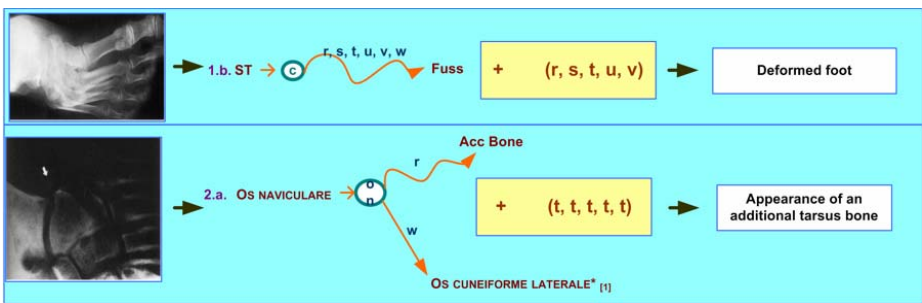


Fig. 5. The automatic understanding of foot bone lesions detected by the UBIAS system

3 Conclusion

To summarise the above discussions, it can be said that the one type of foot projection described above became the basis for introducing definitions of grammars and formal descriptions that are to support the in-depth analysis, interpretation and the semantic description of analysed data in the image form. The analysis and reasoning conducted were not just about the simple identification of the disease, but mainly about identifying the type of that pathology, understanding it and presenting the semantics contained in the image. The attempt at the automatic understanding of the analysed data

by the UBIAS system was made using the cognitive analysis and interpretation of selected medical images depicting various types of foot bone deformations. The depth of the analysis is proven by using not just one, but also by two other foot projections [8]. Identifying very robust formalisms of the linguistic description and the meaning analysis of data makes it possible to conduct a complete analysis of foot bone images, comprising the semantic reasoning and the indication of the specific type of a pathology, but can also be used to indicate specific therapeutic recommendations in the diagnostic/treatment process conducted by a specialist physician (taking into consideration additional information from the patient's medical history).

The described research has demonstrated that an appropriately built image grammar enables the conduct of precise analysis and the description of medical images from which important semantic information can be gained on the nature of processes and pathological lesions. It is worth noting that the results described in this paper have been obtained following the cognitive process, simulating an experts' method of thinking: if one observes a deformation of the organ shown by the medical image used, then one tries to understand the pathological process that was the reason for the appearance of deformations found.

Acknowledgement. This work has been partly supported by the Polish Ministry of Science and Higher Education under Grant No. N519 007 32/0978.

References

1. Burgener, F.A., Meyers, S.P., Tan, R.K., Zaunbauer, W.: *Differential Diagnosis in Magnetic Resonance Imaging*. Thieme (2002)
2. Chan, C.S., Liu, H.: Fuzzy Qualitative Human Motion Analysis. *IEEE Trans. on Fuzzy Systems* 17(3), 1–12 (2009)
3. Liu, H.: A Fuzzy Qualitative Framework for Connecting Robot Qualitative and Quantitative Representations. *IEEE Trans. on Fuzzy Systems* 16(6), 1522–1530 (2008)
4. Meyer-Baese, A.: *Pattern Recognition in Medical Imaging*. Elsevier, Amsterdam (2003)
5. Meystel, A.M., Albus, J.S.: *Intelligent Systems – Architecture, Design, and Control*. John Wiley & Sons, Inc., Chichester (2002)
6. Ogiela, L.: Cognitive Understanding Based Image Analysis Systems (UBIAS) of the Diagnostic Type. In: *IEEE International Workshop on Imaging Systems and Techniques – IST*, Krakow, Poland, May 4-5 (2007)
7. Ogiela, L.: Cognitive Systems for Medical Pattern Understanding and Diagnosis. In: Lovrek, I., Howlett, R.J., Jain, L.C. (eds.) *KES 2008, Part I. LNCS (LNAI)*, vol. 5177, pp. 394–400. Springer, Heidelberg (2008)
8. Tadeusiewicz, R., Ogiela, M.R.: *Medical Image Understanding Technology*. Springer, Heidelberg (2004)
9. Wang, Y.: The Theoretical Framework and Cognitive Process of Learning. In: *Proc. 6th International Conference on Cognitive Informatics (ICCI 2007)*, pp. 470–479. IEEE CS Press, Lake Tahoe (2008)

A New Selective Confidence Measure–Based Approach for Stereo Matching

Nizar Fakhfakh^{1,2}, Louahdi Khoudour¹, El-Miloudi El-Koursi¹, Jacques Jacot²,
and Alain Dufaux²

¹French National Institute for Transport and Safety Research (INRETS)
20, rue Elisée Reclus 59666 Villeneuve d'Ascq, France

{nizar.fakhfakh, louahdi.khoudour, el-miloudi.el-koursi}@inrets.fr

²Ecole Polytechnique Fédérale de Lausanne (EPFL)

CH-1015 Lausanne, Switzerland

{nizar.fakhfakh, jacques.jacot, alain.dufaux}@epfl.ch

Abstract. Achieving an accurate disparity map in a reasonable processing time is a real challenge in the stereovision field. For this purpose, we propose in this paper an original approach which aims to accelerate matching time while keeping a very good matching accuracy. The proposed method allows us to shift from a dense to a sparse disparity map. Firstly, we have computed scores for all pairs of pixels using a new dissimilarity function recently developed. Then, by applying a confidence measure on each pair of pixels, we keep only couples of pixels having a high confidence measure which is computed relying on a set of new local parameters.

Keywords: Correlation function, Confidence measure, Disparity, Stereo.

1 Introduction

Achieving an accurate disparity map in a reasonable processing time is a real challenge in the stereovision field. For this purpose, we propose in this paper an original approach which aims to accelerate matching time while keeping a very good matching accuracy. The proposed method allows us to shift from a dense to a sparse disparity map.

In dense stereovision, several well-known stereo algorithms compute an initial disparity map from a pair of images under a known camera configuration. These algorithms are based loosely on local methods, such as window correlation, which take into account only neighborhood points of the pixel to be matched. The disparity map obtained has a lot of noise and erroneous values. This noise concerns mostly the pixels belonging to occluded or textureless image regions. An iterative process is then applied to the initial disparity map in order to improve it. These methods use global primitives. Cost-relaxation approaches, which were invented by Marr and Poggio [2] and which are picked up again by Brockers [3], belong to this family. Some research has used a graph-based method [5] and color segmentation based stereo methods [4] which belong to what is called “global approaches”. Other approaches have been proposed: they are based on a probabilistic framework optimization, such as expectation-maximization [7] and belief propagation [6, 11]. These methods aim to obtain high-quality and accurate

results, but are very expensive in terms of processing time. It is a real challenge to evaluate stereo methods in the case of noise, depth discontinuity, occlusions and non-textured image regions.

Besides, some mixture of local methods is first used to obtain an estimated disparity map which is improved in a second step by global methods. On the basis of local methods, a pixel on the left-hand image is evaluated with candidate pixels on the corresponding right-hand image. Some research carried out by Yoon [8] applies window-based methods in the Lab color space and are coupled with an adaptive window [10] which tries to find an optimal support window for each pixel. These techniques assume that the neighborhood of a pixel to be matched presents homogeneity in terms of disparity values. In other words, all the pixels in the given correlation window must have very similar disparities.

We present, in this paragraph, an overview of our algorithm which allows us to improve the accuracy of disparity maps. The algorithm can be divided into three parts: Initialization disparity map, pixel classification and disparity allocation. In the first step, we compute the correlation volume. Thus, we take advantage of local methods by applying a new color window correlation to build the correlation volume. It is called Weighted Average Color Difference (WACD) [1]. Firstly, this dissimilarity function is applied to all the pixels in stereo images using a non-adaptive square correlation window. This local method allows an initial appraisal of the disparity map. The second part of our algorithm aims to classify more accurately the matched pixels. Stereo matching accuracy may be affected by various factors including feature descriptors, similarity measures and matching approach. We assume that depth, generally derived from disparity, varies smoothly within color homogeneity in a given region. Depth discontinuities coincide generally with color boundaries or edges. Under the previous constraints, all matched pixels are classified into three categories according to their location: well-matched, badly-matched pixels and unclassified pixels. A new classification method based on a confidence measure approach is applied in this context. This confidence measure is computed for all matched pixels and is based on a set of local parameters referring to scores obtained from a new dissimilarity function. This method and the associated parameters are detailed later in this paper. This work is based on the principle of relaxation and the belief propagation theory which are based on global criteria. However, the difference is that we consider only candidate pixels to evaluate the matched pair.

In this paper, we propose in section 2 an overview of our framework for selected matching pixels and the confidence measure approach which is applied for each matched pixel. The experimental results shown in section 3 demonstrate the advantage and the originality of our approach. We then discuss the proposed method and conclude the paper in section 4.

2 Local Confidence Measure Estimation Theory

In our approach, in order to reduce the processing time and to deal with the problems of ambiguity in the matching process, the correlation function used to evaluate each stereo pair will only be applied on high color variation regions of the images.

The main idea of our approach is to compute a confidence measure for every matched pixel. Furthermore, confidence measure ψ can be seen as a matching probability for pixel P_l (a pixel P in the left-hand image) with pixel P_r (a pixel P in the right-hand image), given some parameters. The way in which confidence measures are calculated is provided by equation 1.

$$\psi(P_l^{i,j}, P_r^{i,v}) = P(P_l^{i,j} / P_r^{i,v}, N, \min, \sigma, \omega) \quad (1)$$

The confidence measure with its parameters is given by equation 2:

$$\psi(P_l^{i,j}, P_r^{i,v}) = \left(1 - \frac{\min}{\omega}\right)^{N^2 \cdot \log(\sigma)} \quad (2)$$

The following subsections provide details on the implementation of the parameters included in equations 1 and 2.

2.1 The Best Correlation Score: \min

The output of the dissimilarity function is a measure representing the degree of similarity between two pixels. Then, the candidate pixels are ranked in increasing order according to their corresponding scores. The couple of pixels that has the minimum score is considered as the best-matched pixels. The lower the score, the better the matching. The nearer the minimum score to zero, the greater the chance of candidate pixel being the right correspondent.

2.2 The Number of Potential Candidate Pixels: N

This parameter represents the number of potential candidate pixels having similar scores. N has a big influence because it reflects the behavior of the dissimilarity function. When the value of N is quite large, that means the first potential candidate pixel is located in a uniform color region of the frame.

The lower the value of N , the fewer the potential candidate pixels. In the case where there are a few candidates, the chosen candidate pixel has a greater chance of being the right correspondent. Indeed, the pixel to be matched belongs to a region with high variation of color components.

While establishing the relationship between N and \min values, with a very small value of N and a minimum score \min , near to zero for instance, the pixel to be matched probably belongs to a region of high color variation.

2.3 The Disparity Variation of N Pixels: σ

A disparity value is obtained for each candidate pixel. For the N potential candidate pixels, we compute standard deviation σ on the N disparity values. A small σ means that the N considered pixels are neighbors. In this case the true candidate pixel should belong to a particular region of the frame, such as edge, transition point. Therefore, it increases the confidence measure. A large σ means that the N candidate pixels taken into account are situated in a uniform color region.

2.4 The Gap Value: ω

This parameter represents the difference between the N^{th} and $(N + 1)^{th}$ scores given with the dissimilarity function used. It is introduced to adjust the impact of the *min* score.

To ensure that this function gives a value between 0 and 1, some constraints are introduced. The *min* parameter must not be higher than the ω one. If so, parameter ω is forced to $min + 1$. However, the $\log(\sigma)$ term is used instead of σ alone. It has a big influence on the confidence measure in the case of high values of σ , and it is indifferent otherwise. This leads to reducing the impact of high values of σ and to obtaining coherent confidence measures.

The number N of potential candidate pixels is deduced from the k scores obtained from each candidate pixel using the dissimilarity function previously presented. The main idea is to detect major differences between successive scores. These differences are called main gaps. Let f denote a function which represents all scores given by the dissimilarity function in increasing order. Then, we apply the average rate growth to the f function. This second function can be denoted by η and can be seen as the ratio of the difference between a given score and the first score, and the difference between their ranks. This function is defined in equation 3.

$$\eta(x_m) = \frac{f_{x_m}^{i,j} - f_{x_1}^{i,j}}{x_m - x_1} \quad m = 1 \dots k \quad (3)$$

where $f_{x_m}^{i,j}$ is the m^{th} of k score of the (i, j) coordinate pixel and x_m is the rank of the corresponding score.

$$\xi(x_m) = \frac{\eta_{x_m}^{i,j} - \eta_{x_{m-1}}^{i,j}}{m^2} \quad m = 1 \dots k \quad (4)$$

The previous function is used to characterize jump scores and is applied only in the case where $(\eta_{x_m}^{i,j} - \eta_{x_{m-1}}^{i,j})$ is a positive value. We have introduced parameter m in order to penalize candidate pixels according to their rank. The number of potential candidate pixels is given by formula 5.

$$N = \underset{m}{\text{Argmax}} \xi(x_m) \quad (5)$$

3 Experimental Results

In this section, we describe the evaluation of the performances of the proposed approach thanks to images with ground truth. Well-known conventional stereo images with available ground truths are employed to test the relevance of the accuracy of our algorithm [9]. In this evaluation, four pairs of stereo images are used: Cones, Teddy, Venus and Sawtooth. As a first step, the disparity map is initialized by applying the dissimilarity function proposed in [1]. This provides a first visual rendering of the disparity map for Cones (Figure 2).

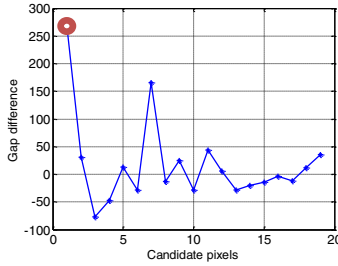


Fig. 1. The number N of potential candidate pixels is the rank of global maximum of ξ function

The dissimilarity function used has the particularity of matching well most pixels in regions having high variation of color components. As shown in Figure 2.d, Pixels belonging to uniform color regions or depth discontinuity regions are badly-matched. This can be explained by the presence of several potential candidate pixels for the given pixel to be matched. Therefore, this increases the error of the matching task.

As a second step, and in order to quantify and to automatically distinguish well-matched pixels from badly-matched pixels, we have exploited the confidence measure method described in section 2. Each couple of matched pixels is evaluated and belongs to one of the following three categories for a given confidence measure threshold T :

- *Well-matched pixels* having a confidence measure higher than T ,
- *Badly-matched pixels* having a confidence measure higher than T and an erroneous disparity values referring to the ground truth image.
- *Unclassified pixels* having a confidence measure lower than T .

The disparity map obtained with an optimal window size (15x15) defined experimentally is taken into consideration. All matched pixels are classified into the three previous categories. Global results are shown in Figure 3.

According to Figure 2.d, the badly-matched pixels represent either occluded pixels or pixels belonging to uniform regions in terms of color. Firstly, in order to reduce the impact of error matching, only pixels belonging to regions of high color variation are considered.

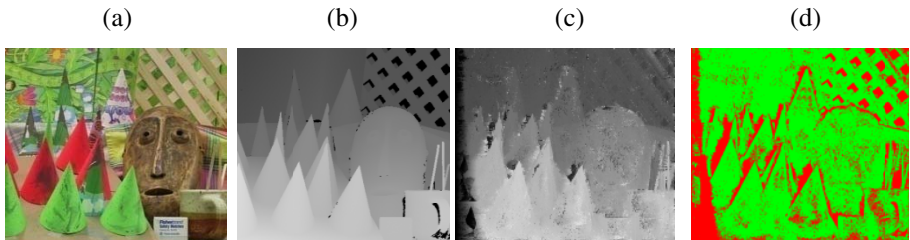


Fig. 2. Left-hand “Cones” image and output images. (a) Left-hand images used for evaluation (Cones) (b) ground truth disparity (c) Disparity map obtained with WACD dissimilarity function application only (d) Well- and badly-matched pixels: Green color for well-matched pixels and red color for badly-matched pixels.

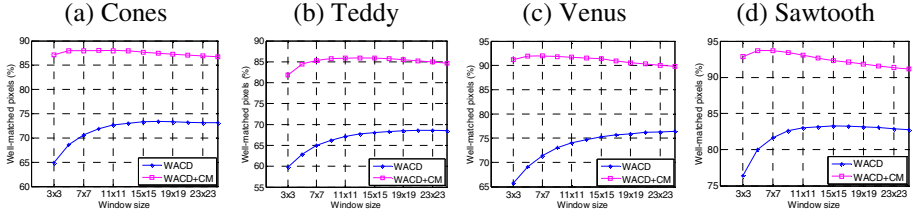


Fig. 3. Well-matched pixels rate using only WACD correlation function (blue curves) and introducing both confidence measure approach and homogeneous color regions elimination (pink color)

In order to reduce the impact of error matching and to have a high gap between the rate of well-matched and badly-matched pixels, only pixels belonging to regions of high color variation are considered. A pixel will be matched only if the sum of all color component variations of neighborhoods pixels belonging to a support window is higher than a threshold. This threshold is computed based on dynamics of color in the image.

In figure 3, four diagrams are used to illustrate the matching performances using the WACD correlation function for the lower curves, and using the same function with introduction of confidence measure for the upper curves. We can notice that the use of the confidence measure applied to matched pixels allows us to improve the matching rates significantly.

Thus, for a correlation window of 15x15 pixels and a 96% confidence measure, the good matching rate passes from:

- 73% to 88% for cones,
- 67% to 86% for Teddy,
- 74% to 92% for Venus,
- 83% to 93% for Sawtooth.

It is to be noted that the good matching rates, with the use of the confidence measure, concern only pixels belonging to high color variation zones (around 80% on average for the four pairs of images tested). The other pixels, those belonging to uniform zones in terms of color, are assessed later. In Figure 4, the different steps of our approach are illustrated. We have extracted a single cone to illustrate visually the improvement for each step. In figure 4.b, for the extracted cone, we can notice that the number of badly-matched pixels (in red) is quite high. These pixels are located in uniform regions of the cone. The application of the confidence measure (passage from 4.b to 4.c) leads to an important decrease in the number of badly-matched pixels. In fact, in figure 4.c we can notice that the number of red pixels has reduced. The white pixels in figure 4.c are the ones with a confidence measure lower than the given threshold (96%). In figure 4.d the unclassified pixels belonging to homogeneous regions in terms of color are identified and marked in white.

Finally, in our method, we have only considered pixels having a high confidence measure. The disparity values of remaining pixels will be assessed later. For this purpose and in order to update the disparity for all the pixels, several directions could be

followed. For instance, it is reasonable to consider that pixels belonging to very uniform regions will have very close disparity values. In this case, by using a segmentation algorithm [12], it will be easy to locate these uniform regions and to allocate them disparity values. Then, statistical methods could be applied to estimate disparity values for unclassified pixels: modal disparity class for a given uniform region, statistical clustering.

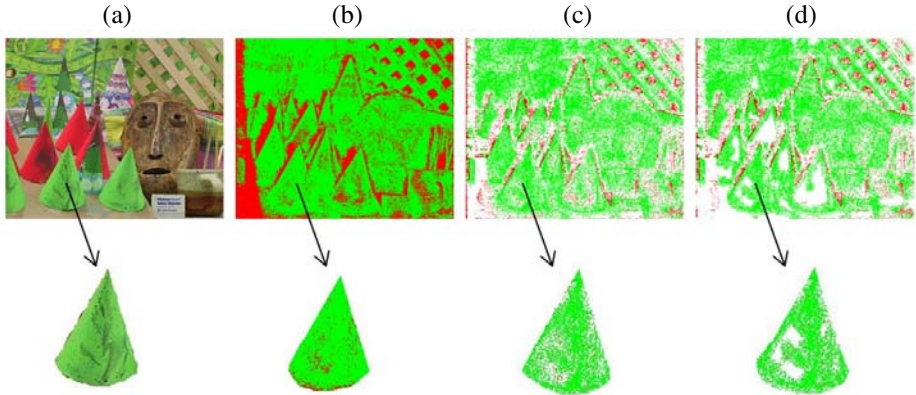


Fig. 4. (a) Left-hand Cones image (b) Well-matched and badly-matched pixels with WACD dissimilarity function application only (c) Well-matched in green, badly-matched in red and unclassified pixels in white for a confidence measure of 96% (d) After elimination of pixels belonging to regions having a low color variation.

4 Conclusion and Perspectives

We have proposed a new method to obtain an accurate disparity map. We have also applied a local matching method to initialize the disparity value for each pixel. We have then introduced new local parameters in order to compute a confidence measure for each matched pixel. The main contribution of our approach is twofold. On the one hand, new parameters introduced above can obtain important information for occluded and uniform region detection. On the other hand, unclassified pixel disparities can be updated in a post-processing step in order to obtain a more accurate disparity map. In fact, our matching algorithms deal well with specific regions of the images: texture, transition points, and high color variations. By working in two steps, that is to say, setting up a first dense disparity map followed by a refinement of it, based on confidence measure theory, allows us to take into account the particularities of the images.

Our approach is highly dependent on the dissimilarity function used for computing the score of all matched pixels. However, a more extensive study will be carried out in order to enhance the dissimilarity function used. The rate of well-matched pixels will, therefore, be improved. The results are encouraging in terms of processing time which is compatible for a real-time implementation. The promising result obtained allows us to follow this track.

References

1. Fakhfakh, N., Khoudour, L., El-Koursi, M.: Mise en Correspondance Stéréoscopique d'Images Couleur pour la Détection d'Objets Obstruant la Voie aux Passages à Niveau. In: TELECOM 2009 & 6ème JFMMA, Agadir, Maroc, p. 206 (4 pages) (2009)
2. Marr, D., Poggio, T.: Cooperative Computation of Stereo Disparity. *American Association for the Advancement of Science* 194(4262), 283–287 (1976)
3. Brockers, R., Hund, M., Mertsching, B.: Stereo Vision using Cost-Relaxation with 3D Support Regions. In: ICVNZ, New Zealand (2005)
4. Taguchi, Y., Wilburn, B., Zitnick, C.L.: Stereo Reconstruction with Mixed Pixels using Adaptive Over-Segmentation. In: CVPR, pp. 1–8, Anchorage, Alaska (2008)
5. Foggia, P., Jolion, J.M., Limongiello, A., Vento, M.: Stereo Vision for Obstacle Detection: A Graph-Based Approach. In: Escolano, F., Vento, M. (eds.) GBRPR 2007. LNCS, vol. 4538, pp. 37–48. Springer, Heidelberg (2007)
6. Lee, C., Ho, Y.: Disparity Estimation using Belief Propagation for View Interpolation. In: ITC-CSCC, Japan, pp. 21–24 (2008)
7. Xiong, W.H., Chung, S., Jia, J.: Fractional Stereo Matching Using Expectation-Maximization. *IEEE TPAMI* 31(3), 428–443 (2008)
8. Yoon, K.J., Kweon, S.: Adaptive Support-Weight Approach for Correspondence Search. *IEEE TPAMI* 28(4) (2006)
9. Scharstein, D., Szeliski, R.: Middlebury stereo vision research page, <http://vision.middlebury.edu/stereo/>
10. Veksler, O.: Fast Variable Window for Stereo Correspondence using Integral Image. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition, Madison, Wisconsin, vol. 1, pp. 556–561 (2003)
11. Sun, J., Zheng, N.-N., Shum, H.Y.: Stereo Matching using Belief Propagation. *IEEE TPAMI* 25(7) (2003)
12. Klaus, A., Sormann, M., Karner, K.: Segment-Based Stereo Matching using Belief Propagation and a Self-Adapting Dissimilarity Measure. In: ICPR, pp. 15–18 (2006)

Image Content Analysis for Cardiac 3D Visualizations

Mirosław Trzupek, Marek R. Ogiela, and Ryszard Tadeusiewicz

AGH University of Science and Technology, Institute of Automatics,
30 Mickiewicza Ave, 30-059 Krakow, Poland
{mtrzupek, mogiela, rtad}@agh.edu.pl

Abstract. The problem tackled in this work is the semantic interpretation and an attempt at computer automatic understanding of a 3D structure of spatially visualised coronary vessels with the use of AI graph-based linguistic formalisms. At the stage of the initial analysis, it was found that the problem is subject to numerous important limitations. These limitations result, among other things from the serious obstacles encountered in the development of a universal standard, defining the model shape of the healthy or diseased organ that could possibly undergo typical recognition. Due to this difficulties a decision was made to apply the methods of automatic image understanding for the interpretation of the images considered, which consequently leads to their semantic descriptions. For this purpose the linguistic approach was applied.

1 Introduction

Today's state-of-the-art methods of image diagnostics provide abundant and varied diagnostic and research material, as 3D visualisations are available for practically all human body structures and organs. They can illustrate the pathological changes to a greater degree and more completely, having however assumed that the physicians interpreting the image understand well, what they see and what it means. Clinical practice shows that, despite the frequent use of major computer-assisted diagnostic equipment, the obtained visualisations of the spatial reconstructions of specific body parts, later undergo only a rough qualitative assessment by the physician diagnostician, with no in-depth semantic analysis. Moreover, often a new method of medical visualisation does present the essence of a medical problem, but the physician watching such a new visualisation is unable to interpret it properly because, in simple words, there are so many things shown that he or she does not know what to look at. To make matters even worse, the progress in the visualisation technology mentioned still accelerates at greater and greater speeds, which means that by the time today's alumni become experienced masters in the field of interpreting certain state-of-the-art forms of medical visualisation, these forms of visualisation will have become dated and replaced by successive, even finer, yet again not in line with the experience acquired by the physicians. This shows that the actual progress in the practical application of 3D visualisations in medicine may depend on the progress in developing smart diagnostics support systems, making use of automatic analysis and understanding of medical 3D images.

2 Semantic Models for Spatial Reconstruction of the Heart's Coronary Vessels

To allow the introduction of linguistic formalisms, several visualisations for various patients during diagnostic examinations of the heart were obtained with a helical CT scanner with 64 detectors. Such visualisations present in a very clear manner all morphologic changes of individual sections of arteries in any plane. The structures made visible in this manner will be described with the use of graph-based grammars – constituents of the graph-based language defining their proper, spatial topology [2, 6, 7, 8]. Spatial reconstructions of the coronary vessel obtained from a spiral CT allows the structures significant for further analysis to be extracted quickly, showing their morphology changes. In practice, this technique turns out to be among the most precise image diagnostic techniques, better in terms of quality and detail of the information provided (e.g. the functional conditions of the heart vascularisation) than other techniques used to study vessel morphology [9], where it also becomes necessary to extract the vessels being studied using advanced segmentation techniques [1, 5]. Contemporary methods of visualization in medicine by means of specialized techniques as CT, provide images with high display resolution and visual quality that enable a precise construction of a 3D solid figure that is an equivalent of a real object [4, 10]. For that reason, the problem of noisy pixels did not exist and on the image data tested there were not such images.

In order to analyze a 3D reconstruction, it becomes necessary to select the appropriate projection showing lesions in vessels in a way that enables them to be analysed on a plane. In the clinical practice, this operation is done manually by the operator, who uses his/her own criteria to select the appropriate projection which shows the coronary vessels including their possible lesions. In our research we have attempted to automate the procedure of finding such a projection by using selected geometric transformations during image processing. Next, to enable a linguistic representation of the spatial reconstructions studied, the coronary vessels shown in them had been subjected to the operation of thinning, referred to as skeletonising. This operation allows us to obtain a skeleton of the arteries under consideration with the thickness of one unit. This skeleton can then be subjected to the operation of labelling, which determines the start and end points of main and surrounding branches of coronary arteries in it. These points will constitute the peaks of a graph modelling the spatial structure of the coronary vessels of the heart. The next step is labelling them by giving each located informative point the appropriate label from the set of peak labels which unambiguously identify individual coronary arteries forming parts of the structure analysed. In the case of terminal points (leaves of a graph modelling the coronary vascularisation), the set of peak labels comprises abbreviated names of arteries found in coronary vascularisation. They have been defined as follows:

For the left coronary artery: LCA - left coronary artery, LAD - anterior interventricular branch (left anterior descending), CX - circumflex branch, L - lateral branch, LM - left marginal branch.

For the right coronary artery: RCA - right coronary artery, A - atrial branch, RM - right marginal branch, PI - posterior interventricular branch, RP - right posterolateral branch.

If a given informative point is a branching point, then the peak will be labelled with the concatenation of names of the peak labels of arteries which begin at this point. An exception here is the main branching of the left LCA artery into the circumflex branch CX and the main branching of the right RCA artery into the marginal branch RM (if the distribution of arteries is balanced) or into the atrial branch A (if the right artery is dominant). This is due to the fact that the above nodes together with the starting point - the root of the graph ST - determine the sections along which lesions are searched for in both the left and the right coronary artery. This way, all initial and final points of coronary vessels as well as all points where main vessels branch or change into lower level vessels have been determined and labelled as appropriate. After this operation, the coronary vascularisation tree is divided into sections which constitute the edges of a graph modelling the examined coronary arteries. This makes it possible to formulate a description in the form of edge labels which determine the mutual spatial relations between the primary components, i.e. between subsequent arteries shown in the analysed image. These labels have been identified according to the following system. Mutual spatial relations that may occur between elements of the vascular structure represented by a graph are described by the set of edge labels presented in Fig. 1.

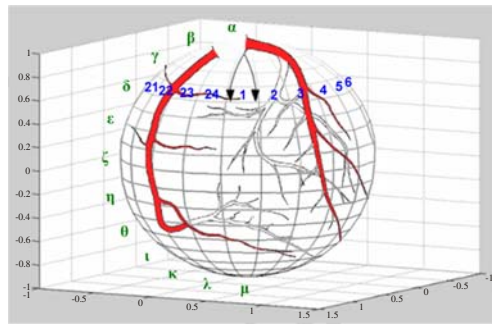


Fig. 1. A set of edge labels of graphs describing mutual spatial relations between individual coronary arteries

The elements of this set have been defined by introducing the appropriate spatial relations: vertical, defined by the set of labels $\{\alpha, \beta, \dots, \mu\}$ and horizontal, defined by the set of labels $\{1, 2, \dots, 24\}$ on a hypothetical sphere surrounding the heart muscle. These labels designate individual final intervals, each of which has the angular spread of 15° . Then, depending on the location, terminal edge labels are assigned to all branches identified by the beginnings and ends of the appropriate sections of coronary arteries. The presented methodology draws upon the method of determining the location of a point on the surface of our planet in the system of geographic coordinates, where a similar cartographic projection is used to make topographic maps. This representation of mutual spatial relationships between the analysed arteries yields a convenient access to and a unanimous description of all elements of the vascular structure. At subsequent analysis stages, this description will be correctly formalised using ETPL(k) graph grammars, supporting the search for stenoses in the lumen of arteries forming parts of the coronary vascularisation.

As the structure of coronary vascularisation may be characterised by three different types of artery distribution over the heart surface, in the following part we will propose a grammar for the right dominance artery distribution. The right dominance artery distribution is present in on average 20–24% of cases, with a variety of intermediate forms possible. In that case, the circumflex branch of the left coronary artery is highly retarded while the right coronary artery is highly developed and is the main supplier of blood to the posterior surface of the left chamber.

Before we define the representation of the analysed image in the form of IE graphs, we have to introduce the following order relationship in the set of Γ edge labels (shown in Fig. 1): $1 \leq 2 \leq 3 \leq \dots \leq 24$ and $\alpha \leq \beta \leq \gamma \leq \dots \leq \mu$.

This way, we index all peaks according to the \leq relationship in the set of edge labels which connect the main peak marked 1 to the adjacent peaks and we index in the ascending order ($i = 2, 3, \dots, n$). This gives us IE graphs for the right and the left coronary arteries, respectively, presented in Fig. 2. When graphs shown in Fig. 2 are represented by their characteristic descriptions, they look as presented in Table 1 and 2.

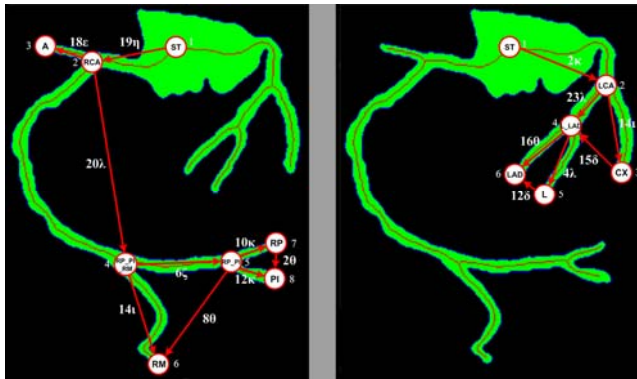


Fig. 2. The representation of the right (A) and the left (B) coronary artery using IE graphs

Table 1. Description for the right coronary artery

ST ₁	RCA ₂	A ₃	RP_PI_RM ₄	RP_PI ₅	RM ₆	RP ₇	PI ₈
1	2	–	2	2	–	1	–
19η	18ε 20λ	–	6ζ 14τ	10κ 12κ	–	2θ	–
2	3 4	–	5 6	7 8	–	8	–

Table 2. Description for the left coronary artery

ST ₁	LCA ₂	CX ₃	L_LAD ₄	L ₅	LAD ₆
1	2	1	2	1	–
2κ	14τ 23μ	15δ	4λ 16θ	12δ	–
2	3 4	4	5 6	6	–

The graph structure created in this way will form elements of a graph language defining the spatial topology of the heart muscle vascularisation including its possible morphological changes. Formulating a linguistic description for the purpose of determining the semantics of the lesions searched for and identifying (locating) pathological stenoses will, *inter alia*, support the computer analysis of the structure obtained in order to automatically detect the number of stenoses, their location, type (concentric or eccentric) and extent.

For IE graphs defined as above, in order to locate the place where stenoses occur in the case of a balanced artery distribution, the graph grammar may take the following form.

For the right coronary artery: $G_p = (\Sigma, \Delta, \Gamma, P, Z)$

$\Sigma = \{ST, RCA, A, RP_PI_RM, RP_PI, RM, RP, PI, C_Right, C_Right_post_marg, C_Right_post_int\}$

$\Delta = \{ST, RCA, A, RP_PI_RM, RP_PI, RM, RP, PI\}$, $\Gamma = \{19\eta, 18\epsilon, 20\lambda, 6\xi, 14\tau, 10\kappa, 12\kappa, 2\theta, 8\theta\}$, Z is the start graph shown in Fig. 3.

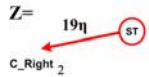


Fig. 3. Start graph Z for grammar G_p

P is the set of productions shown in Fig. 4.

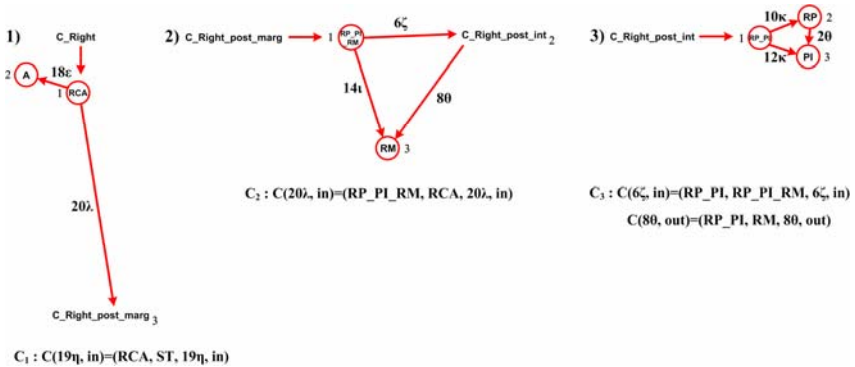


Fig. 4. Set of productions for grammar G_p

For the left coronary artery: $G_L = (\Sigma, \Delta, \Gamma, P, Z)$

$\Sigma = \{ST, LCA, CX, L_LAD, L, LAD, C_Left, C_Left_ant\}$, $\Delta = \{ST, LCA, CX, L_LAD, L, LAD\}$, $\Gamma = \{2\kappa, 14\tau, 23\lambda, 4\lambda, 16\theta, 15\delta, 12\delta\}$, Z is the start graph shown in Fig. 5.

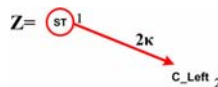


Fig. 5. Start graph Z for grammar G_L

P is the set of productions shown in Fig. 6.

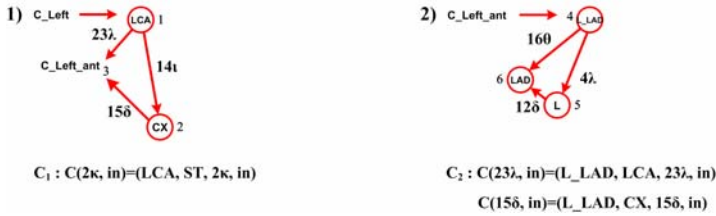


Fig. 6. Set of productions for grammar G_L

This way, we have defined a mechanism in the form of ETPL(k) graph grammars which create a certain linguistic representation of each analyzed image in the form of an IE graph. The set of all representations of images generated by this grammar is treated as a certain language. Consequently, we can build a syntax analyzer based on the proposed graph grammar which will recognize elements of this language. The syntax analyzer is the proper program which will recognize the changes looked for in the lumen of coronary arteries. It is, of course, the most important and difficult (from the implementation point of view) part in the whole task especially for the graph and tree grammars [3, 6, 7, 8]. The difficulties with implementing syntax analyzers based on graph grammars are due to the lack of ready grammar compilers, like those available for context-free grammars. As a result, it becomes necessary to independently execute syntax analysis procedures for the proposed grammars. The authors have used in their research a prototype computer system which allows graph descriptions to be generated, and then used to create their corresponding parsers. This system was developed as part of one of the scientific projects dedicated to problems of automating grammatical reasoning processes described in publication [3]. The entire such syntax and semantic analysis can be completed in the multinomial time, characterized by the complexity of $O(n^2)$ both for strictly defined patterns, and also for fuzzy patterns, as the grammars described can be extended to probabilistic forms. This means that they constitute extremely efficient analyzers which enable a fast structural verification of the pattern described and its compliance with the grammar rules introduced. Theoretical aspects of conducting the syntax analysis and constructing parsers for ETPL(k) grammars are described in [6, 8]. This property is convenient, particularly when it is necessary to analyze patterns not taken into account previously to reason out grammar rules.

3 Summary

The image set of recognition data, which has been used in order to determine in percentage figures the efficiency of a correct recognition of the size of stenoses in coronary arteries, included 16 different spatial reconstructions obtained for patients with heart disease (mostly ischemia). In this set, we considered image sequences of patients previously analysed at the stage of the grammar construction and the recognising analyser. In order to avoid analysing identical reconstructions we selected

separate images occurring after slight positions rotation (different projection) the ones used originally (from spatial helical CT scans). The remaining images in the test data have been obtained for a new group of patients. The objective of an analysis of these data was to determine in percentage the efficiency of the correct recognition of artery stenosis and to determine their size with the use of the grammar introduced.

The recognition of such stenoses, including the determination of their locations, lumens of the artery, and the types (concentric or eccentric), was conducted in such a way that while reasoning out the grammar for the graph representation of the coronary vascularisation, particular edges of the graph determined the actual beginnings and ends of particular sections of coronary arteries. During the grammar reasoning and the course of the transform of embedding graph representations on the actual images, the corresponding sections of arteries were analysed with regard to the presence of potential stenoses in them. The method of this analysis also consisted in applying a context-free sequential grammar to detect stenoses in 2D coronarography images. Such a grammar has been defined in publication [6, 7, 8], while an example of a diagnosis of a stenosis applying this grammar to a coronarography image is presented in Fig. 7. Applying such a grammar to analyse particular sections of arteries in the obtained spatial reconstructions turned out to be quite effective, as it allowed the unanimous location of the lesion present together with defining their size and type.

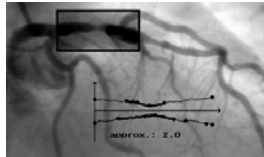


Fig. 7. An example result of coronary artery stenosis detection using context-free grammar

On the image data tested, the efficiency of recognition amounted to 85%. The value of the efficiency of recognition is determined by the percentage fraction of the accurately recognized and measured vessel stenoses compared to the number of all images analyzed in the test. The recognition itself meant locating and defining the type of stenosis, e.g. concentric or eccentric. In order to estimate the correct value of the narrowing degree, obtained with the use of grammars semantic actions, several comparative values from syngo Vessel View [11] (integrated with the HeartView CI package) clinical application were used. Such system is often used in clinical practice with the SOMATOM Sensation Cardiac 64, which was equipped with vessel segmentation routines allowing for measurement stenosis quantification. However in order to confirm or deny the type of stenosis (concentric or eccentric), on the analysed image, it was decided to perform visual estimation. In fact such option in mentioned software has not been implemented. Of course in various places many specialized programs are used, where 3D medical objects can be displayed. For example popular in medical community is computer program 3D-DOCTOR [12].

It is worth pointing out that this work presents a new approach to the modelling and the meaning analysis of coronary vessel reconstructions. The analysis of such structures constitutes an important element of medical information technology methods. It was for this purpose that the authors proposed an approach in which coronary

vascularisation structures are analysed using graph image languages. Such methods extend the authors' previous research on using mathematical linguistics formalisms for the semantic analysis of various medical images. This publication marks the first time that the opportunities for using methods based on parsing graph grammars to detect lesions in medical images are presented.

The approach to cognitive interpretation of spatial reconstructions of coronary vessels presented in this work is a significant innovation among the existing approaches to the computer-assisted medical diagnostics. Such methods are of profound significance as they allow not only for the discovery of pathologic changes but also for supporting description of their semantics, which – in the case of medical diagnostic images – may lead to computer-assisted ‘understanding’ of their medical significance and in future also to fine-tune the optimum therapeutic options.

Acknowledgement. This work has been supported by the Ministry of Science and Higher Education, Republic of Poland, under project number N519 007 32/0978.

References

1. Hoffman, K.R., Sen, A., Lan, L.: A system for determination of 3D vessel tree centerlines from biplane images. *The International Journal of Cardiac Imaging* 16, 315–330 (2000)
2. Meyer-Baese, A.: *Pattern Recognition in Medical Imaging*. Elsevier-Academic Press, San-Diego (2003)
3. Ogiela, M.R., Tadeusiewicz, R.: *Modern Computational Intelligence Methods for the Interpretation of Medical Images*. Springer, Heidelberg (2008)
4. Lewandowski, P., Tomczyk, A., Szczepaniak, P.S.: Visualization of 3-D Objects in Medicine - Selected Technical Aspects for Physicians. *Journal of Medical Informatics and Technologies* 11, 59–67 (2007)
5. Seghers, D., Loeckx, D., Maes, F., Vandermeulen, D., Suetens, P.: Minimal shape and intensity cost path segmentation. *IEEE Trans. on Medical Imaging* 26, 1115–1129 (2007)
6. Skomorowski, M.: *A Syntactic-Statistical Approach to Recognition of Distorted Patterns*. Jagiellonian University, Krakow (2000)
7. Tadeusiewicz, R., Flasiński, M.: *Pattern Recognition*. Warsaw (1991)
8. Tadeusiewicz, R., Ogiela, M.R.: *Medical Image Understanding Technology*. Springer, Berlin (2004)
9. Wild, P.S., Zotz, R.J.: Fragment reconstruction of coronary arteries by transesophageal echocardiography - A method for visualizing coronary arteries with ultrasound. *Circulation* 105, 1579–1584 (2002)
10. Wróbel, K., Porwik, E., Porwik, P.: Three dimensional image projections and its measurement using the vrmI technique. *Journal of Medical Informatics and Technologies* 11, 123–133 (2007)
11. *Get the Entire Picture, SOMATOM Sensation Cardiac 64 Brochure*, Siemens medical (2004)
12. <http://www.ablesw.com/3d-doctor/index.html>

Illogical Adjective Phrase Detection for Computer Conversation

Eriko Yoshimura, Seiji Tsuchiya, Hirokazu Watabe, and Tsukasa Kawaoka

Dept. of Intelligent Information Engineering & Sciences,
Faculty of Science and Engineering, Doshisha University,
Kyo-Tanabe, Kyoto, 610-0394, Japan
{eyoshimura, stsuchiya, watabe, kawaoka}@indy.doshisha.ac.jp

Abstract. We propose an illogical discourse judgment technique using a concept association system with the aim of enabling computer-generated logical discourse. We focused on a relation of nouns and adjective phrases. Then the knowledge structure of how to use nouns and adjective phrases is modeled by arranging the relation in a point of wrongness. Also, this paper proposes a technique for detection relation of nouns and adjective phrases by creating a knowledge model from generation of response sentences. This paper discusses detecting method illogical combinations of words. We showed that this technique was able to very accurately judge illogical usages with 87% accuracy, thus demonstrating the effectiveness of the technique.

Keywords: Illogical phrase, Computer interface human factors, Knowledge engineering, Knowledge representation, Natural languages.

1 Introduction

By enabling machines to engage in smooth discourse with humans, it will be possible to create intelligent robots that can communicate effectively with people. Thus, a great deal of attention is being paid to research on natural language processing for enabling natural discourse. However, research into natural language processing has focused mainly on discourse processing for certain limited objectives and under specific conditions, with emphasis on the superficial formation of sentences [1]. Also, there has been a strong tendency to collect large quantities of response examples to create a knowledge base, although the understanding of such user-generated words is constrained by the size of the constructed knowledge base and the quality of data collected. Because this system does not include common sense and is unable to understand information from the surrounding environment or discourse partner, discourses tend to follow a fixed pattern, resulting in illogical responses.

If a computer has typical responses, and the changeable parts of sentences can be changed by association, more flexible and more various conversations can be done [2]. However, there is a risk that the generation of response sentences by a computer results in a combinations of feeling of wrongness caused by the mechanical combination of words. This paper focused on a relation of nouns and adjective phrases. Then the knowledge structure of how to use nouns and adjective phrases is modeled by arranging the relation in a point of feeling of wrongness. Also, this paper proposes a

technique for detection relation of nouns and adjective phrases by creating a knowledge model from generation of response sentences. This paper discusses detecting method illogical combinations of words. The occurrence of logical responses indicates to the user that the machine has understood the meaning and is able to converse with common sense. It is expected that this ability will enable very free, flexible discourse with machines. In order to show the possession of common sense, we are conducting research on the development of techniques to prevent the generation of illogical discourse. In this paper, “illogical discourse” is used to mean “strange expressions.” These include expressions that make listeners feel that something is wrong and expressions that seem unnatural.

2 Purpose

Humans can immediately identify what is unnatural about the discourse and why it is unnatural. This is because humans possess common sense in relation to words. Thus, in order for computers to be aware that “this is unnatural,” or “this is a strange expression,” they, too, must possess common sense in relation to words. In other words, systems that can deal with such sentences must also be capable of understanding their meaning and responding to them, based on common sense. Especially, we describe the detection method of illogical phrase about percept. This “percept” indicates a sense that can be acquired by stimulation through any of the five percepts (vision, hearing, smell, taste, and touch). The system using this method judge illogical adjective phrase which needs a noun-related percepts. That is to say, like “black apple”, it judge suitability of the relation between a noun and a word expressing the noun. Therefore, in this study, we propose a method for recognizing illogical discourse. Our system is constructed using the Japanese language.

3 Relation between Noun and Adjective Phrase

We arranged the relation between noun and adjective phrase. For example, there is a noun “apple”. “Red, sweet, and round” are description word that we generally associate “Apple” with. These description words are characteristic adjective word on expression of “apple”. A relation between these adjective words and “Apple” is natural. But, a relation between “apple” and “white and black” that expresses a color like “red” is illogical (ex. black apple). Like this, there is adjective phrase that we generally associate a noun with. On the other hand, there is adjective phrase that is right logically although we don't associate a noun with the word. For example, we don't feel illogical for “heavy apple” and “light apple”. These aren't characteristic adjective of “apple”. So we don't associate these words. But “apple” have mass. So the expression of “heavy and light” is right logically and natural about “apple”. These relations are arranged like next 4 groups.

- 1) Characteristic: Adjective expressing characteristic of an object
red apple, yellow banana, round earth, wide sea
- 2) Opposite-characteristic: Adjective of opposite character from characteristic
black apple, black banana, square earth, narrow sea

- 3) Logistic: not characteristic word of an object
black car, red balloon, old book, heavy door
- 4) Anti-logistic: an object can not have the characteristic
square illness, salty sunset, low gloves, bright mud
“Logistic” and “Anti-logistic” are correct logically but we don’t associate generally. On the conversation, we do not feel incongruity about phrase of “Characteristic” and “Logical”. But we feel incongruity about phrases of “Opposite-characteristic” and “Anti-logistic”.

About “Opposite-characteristic” and “Anti-logistic”, for example we considered an expression “square water-melon”. Water-melon is round or ellipse generally, but there is square water-melon that is made for exchange by fitting in a square frame while young. Like expression of “square water-melon”. The expression like this is not use generally on conversation, and that is why it has a value as news. It has an effect to attract us, so it is used novel’s title and message of advertisement. The phrase becomes more attractive because we feel sense of incongruity in the expression. On the detection method proposing this paper, the system detect this expression as illogical, because the purpose of this method is the use to conversation, so in a natural discourse, if someone feels that the discourse is unnatural, then he or she will typically respond by expressing doubt to the discourse partner.

4 Knowledge for Judging Illogical Adjective Phrase

It needs a knowledge structure about general property of objects. For example, “apple” has characteristics as “red, round, sweet”, and has property “colour, shape, taste, smell and weight”, but doesn’t have property “brightness and sound”. The idea of characteristics and property can express effectively by thesaurus structure. Thesaurus [3] is a dictionary where words are semantically classified and generally indicated with a tree structure. The thesaurus has two types: 1) a classification thesaurus with words only on leaf nodes and 2) a hierarchical thesaurus with words on root nodes and intermediate nodes besides leaf.

Using thesaurus structure, parent node property is succeeded child node. Child node and leaf have characteristics adjective. Add, other nodes that do not succeed the property “taste” can not express words “taste” adjective words like “delicious, nasty” (ex. delicious dictionary). Then, a leaf “lemon” is given characteristic adjective “sour”. By this, the knowledge base express that “lemon” can express by “sour” (ex. sour lemon). And, lemon is not given no-usual characteristic for lemon like “spicy, sweet, and salty”. By this, the knowledge base express that it is difficult to express “lemon” in these words. To express the relation of characteristic and property, knowledge base of proposal method is constructed using a relation leaf and node of thesaurus. 680 words used often in daily life are registered as leaf (representative word) and the words are given characteristic adjective words respectively. Moreover, 153 words that group these leaf words is registered as a node and adjective words of general property of the group is given. The image is shown figure 1. This is registered some basic word by human. The proposal method of this paper supplements representative knowledge by human with general knowledge base “Concept Base [4]” and “Degree of Association [4]”. By this mechanism, the proposal method keeps generality in knowledge.

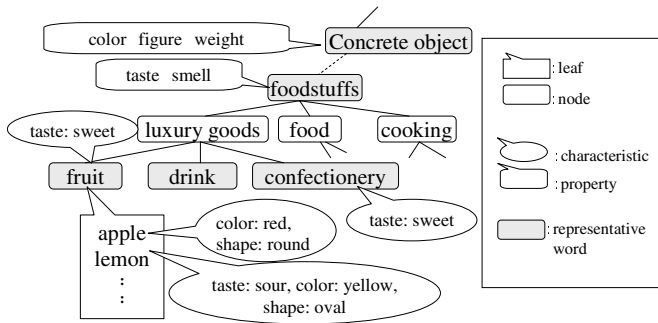


Fig. 1. Knowledge base image for judging illogical adjective phrase

5 Concept Association Mechanism

5.1 Concept Base and Degree of Association

The Concept Association Mechanism incorporates word-to-word relationships as common knowledge. This is a structure for capturing various word relationships. This section describes the Concept Base and the Degree of Association.

The Concept Base is a knowledge base consisting of words (concepts) and word clusters (attributes) that express the meaning of these words. This is automatically constructed from multiple sources, such as Japanese dictionaries and contains approximately 120,000 registered words organized in sets of concepts and attributes. An arbitrary concept, A , is defined as a cluster of paired values, consisting of attribute, a_i , which expresses the meaning and features of the concept, and weight, w_i , which expresses the importance of attribute a_i , in expressing concept A :

$$A = \{(a_1, w_1), (a_2, w_2), \dots, (a_N, w_N)\}$$

Attribute a_i is called the first-order attribute of concept A . In turn, an attribute of a_i (taking a_i as a concept) is called a second-order attribute of concept A .

train	train, 0.36	railroad, 0.10		a_i, w_i	Primary Attributes
	train, 0.36	railroad, 0.10	...	a_{i1}, w_{i1}	
	railroad, 0.10	subway, 0.25	...	a_{i2}, w_{i2}	Secondary Attributes
	:	:	...	:	
	a_{ij}, w_{ij}	a_{2j}, w_{2j}	...	a_{ij}, w_{ij}	

Fig. 2. Example demonstrating the Concept “train” expanded as far as Secondary Attributes

Figure 2 shows the elements of the Concept “train” expanded as far as the Secondary Attributes. The method for calculating the Degree of Association involves developing each concept up to second-order attributes, determining the optimum combination of first-order attributes by a process of calculation using weights, and evaluating the number of these matching attributes.

For Concepts A and B with Primary Attributes a_i and b_j and Weights u_i and v_j , if the numbers of attributes are L and M , respectively ($L \leq M$), the concepts can be expressed as follows:

$$A = \{(a_1, u_1), (a_2, u_2), \dots, (a_L, u_L)\}$$

$$B = \{(b_1, v_1), (b_2, v_2), \dots, (b_M, v_M)\}$$

The Degree of Identity (A, B) between Concepts A and B is defined as follows (the sum of the weights of the various concepts is normalized to 1):

$$I(A, B) = \sum_{a_i=b_j} \min(u_i, v_j) \tag{1}$$

The value of the Degree of Association is a real number between 0 and 1. The higher the number is, the higher the association of the word. Table 1 lists examples of the degree of association.

Table 1. Examples of the degree of association

Concept A	Concept B	Degree of association between A and B
Flower	Cherry blossom	0.208
Flower	Car	0.0008
Car	Bicycle	0.23

5.2 Unknown Word Processing

By using the Concept Base and the Degree of Association, an unknown word that doesn't exist in knowledge base can be processed [5]. Unknown word doesn't exist in knowledge base that is made by human but exist in Concept Base (90000 concepts). Now, it uses as an example of unknown word "Jade"(figure 3).

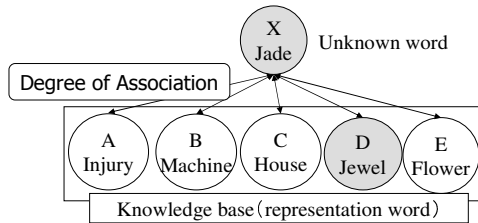


Fig. 3. Example of unknown word processing

First, it calculate the Degree of Association between this unknown word "Jade" and nodes of knowledge base and it link a node with the highest degree of association for "Jade". For "Jade", the node with the highest degree of association is "Jewel". By this, it considered that "Jade" has property of "Jewel". Moreover, attributes of "Jade" is acquired from the Concept base. It considered that the characteristic words of "Jade" are adjective words in these attributes having strong relation for "Jade". By these processes, unknown word "Jade" can have property and characteristic.

6 Illogical Discourse Processing System

For detecting illogical discourse, there are methods of using the database made by statistical value of words and method of using the database made by human. First, using the database made by statistical value of words, it can detect illogical discourse by searching a set of object words. However, if an expression appears even once, it is judged a general expression. Therefore, it can't capture an illogical discourse on purpose to attract public notice; a phrase of ad, a title of novel and so on. Second, using the database made by human, it can detect logical discourse because the human arrange the database logically and generally. However, it is impossible to store data all-inclusive and the data might be different depending on a manufacturer.

The Commonsense Judgment system also associates on the word that doesn't exist in the database made by human. Therefore, using the method described in this report, the covered range can be expanded more than the database only made by human.

Illogical discourse processing involves a judgment component and a response component, but the explanation in this paper focuses on the judgment component.

In order to perform illogical discourse judgments, it is first necessary to extract words from a text to serve as judgment objects. For the object words to appear, a fixed pattern must exist in the text structure. Thus, we created a database of fixed patterns. Then, by performing a text structure analysis of input sentences and determining whether or not these matched the patterns, we extracted judgment object words. We called this database of collected fixed patterns "text structure patterns."

In order to analyze the text structure, we use a meaning understanding system. The system stores input text (single sentences) by dividing them into 6W1H (what, who, whom, why, where, how) + verb frames. Figure 4 shows an example of the meaning understanding system.

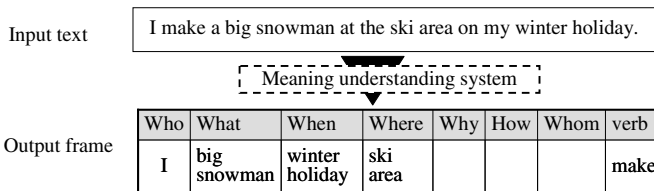


Fig. 4. Example of using a meaning understanding system

When using the meaning understanding system, the combination of frames containing object words has a fixed pattern. For this reason, text structure patterns are selected according to the presence or absence of words in the frames and part-of-speech patterns. The text structure patterns include five patterns for percept. This database also stores the relationships between two extracted words for each pattern.

By comparing the frames resulting from inputting the text into the meaning understanding system with the text structure patterns, it is possible to extract words for use as judgment objects. This method makes it easier to extend the system by adding similar rules, even when introducing new common-sense factors.

The extracted object words are judged for each factor to determine whether they are illogical. With the percept factor, the object words are nouns and descriptive words. For judging their relationship, it needs to have a knowledge structure relating to the common-sense qualities and characteristics of the object nouns. This is Knowledge for judging illogical adjective phrase. By using this knowledge and unknown word processing, this system judges a phrase as illogical. Figure 5 shows a flow chart of an illogical judgment process.

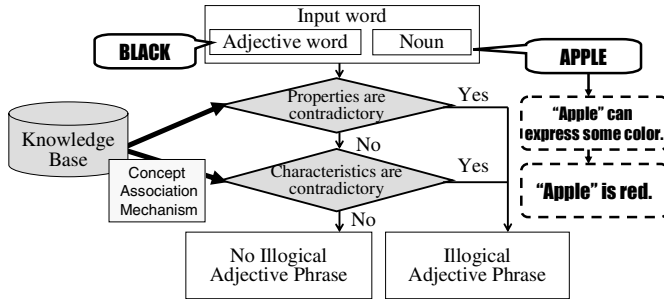


Fig. 5. Flow chart of illogical judgment

7 Evaluation

We evaluated the proposed illogical discourse judgment technique. We manually prepared 100 illogical discourse texts and 100 logical discourse texts. For each of these texts, we evaluated the proportion of correctly classified texts using the illogical discourse judgment technique.

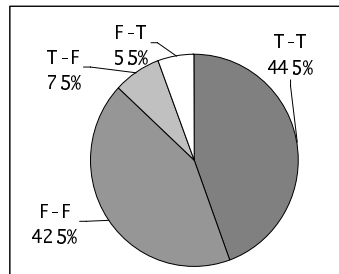


Fig. 6. Evaluation results

Figure 6 shows the evaluation of the illogical discourse judgment technique. “F-F” indicates the judgment of illogical texts as illogical; “T-T” indicates the judgment of logical texts as logical; “F-T” indicates the judgment of illogical texts as logical; and “T-F” indicates the judgment of logical texts as illogical. In this study, we calculate the accuracy as the total of “F-F” and “T-T” results as a proportion of all results.

The accuracy was 87%. Table 2 shows success examples and failure examples using the proposal system. For example, the system can judge that “tasting spicy honey” was judged as illogical, and “drinking cold beer” as logical.

By the lack of some basic data of the knowledge base, the failure occurred. But it is shown that it is effective for almost general words though it was not completely covered. Because the evaluation sentences are gathered by persons who don't see the inside of the system and are different from the system designer. Thus, through this method we showed that this judgment system is effective.

Table 2. Success and Failure example

	Success example		Failure example	
Characteristic	T-T	I ate <u>red strawberries</u> .	T-F	I ate <u>white rice</u> .
Opposite-characteristic	F-F	A <u>dry ice</u> is <u>warm</u> .	F-T	<u>Square tomato</u> was bought.
Logistic	T-T	An <u>old magazine</u> was read.	T-F	I lost my way in a <u>deep forest</u> .
Anti-Logistic	F-F	She use a <u>low purse</u> .	F-T	They go to a <u>round school</u> .

8 Conclusion

In this study, we propose an illogical discourse judgment technique using a concept association system with the aim of enabling computer-generated logical discourse. We created a knowledge structure model for detecting illogical words. Furthermore, using this knowledge structure, we devised an illogical discourse judgment system. Using the method described in this report, we showed that this technique was able to very accurately judge illogical usages with 87% accuracy, thus demonstrating the effectiveness of the technique. By constructing a system capable of handling illogical discourses, machines can demonstrate to users that they possess logic, or common sense, and the capacity to understand discourses, thereby pushing machines one step closer to human-like conversation.

Acknowledgements

This research has been partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (Young Scientists (B), 21700241).

References

1. Weizenbaum, J.: ELIZA - A Computer Program For the Study of Natural Language Communication Between Man and Machine. Communications of the Association For Computing Machinery 9(1), 36–45 (1965)
2. NTT Communication Science Laboratory, NTT Thesaurus, NIHONGOGOITAIKEI (Iwanami Shoten book) (1997)

3. Kojima, K., Watabe, H., Kawaoka, T.: A Method of a Concept-base Construction for an Association System: Deciding Attribute Weights Based on the Degree of Attribute Reliability. *Journal of Natural Language Processing* 9(5), 93–110 (2002)
4. Tsuchiya, S., Watabe, H., Kawaoka, T.: A Sensuous Association Method Using an Association Mechanism for Natural Machine Conversation. In: Gabrys, B., Howlett, R.J., Jain, L.C. (eds.) KES 2006. LNCS (LNAI), vol. 4251, pp. 1002–1010. Springer, Heidelberg (2006)

A Non-sequential Representation of Sequential Data for Churn Prediction

Mark Eastwood and Bogdan Gabrys

Computational Intelligence Research Group,
School of Design, Engineering and Computing,
Bournemouth University
{meastwood, bgabrys}@bournemouth.ac.uk

Abstract. We investigate the length of event sequence giving best predictions when using a continuous HMM approach to churn prediction from sequential data. Motivated by observations that predictions based on only the few most recent events seem to be the most accurate, a non-sequential dataset is constructed from customer event histories by averaging features of the last few events. A simple K-nearest neighbor algorithm on this dataset is found to give significantly improved performance. It is quite intuitive to think that most people will react only to events in the fairly recent past. Events related to telecommunications occurring months or years ago are unlikely to have a large impact on a customer's future behaviour, and these results bear this out. Methods that deal with sequential data also tend to be much more complex than those dealing with simple non-temporal data, giving an added benefit to expressing the recent information in a non-sequential manner.

1 Introduction

In the telecommunications industry, it has been estimated [13] that on average it can cost between 5-8 times more to gain a new customer than it would to keep an existing customer (for example by offering a small incentive). However this incentive is wasted if it is not offered to someone who, in the near future, is likely to churn (that is, to leave the company for a competitor). The high churn rate prevalent in this area means that fairly small improvements in the accuracy of churn prediction can mean significant cost savings. Thus the problem of predicting customer churn is an important one.

It is a very difficult problem. Though we have large quantities of data available, it is limited in that many possible reasons for churn will likely leave no imprint in this data, for example competitor's offers, or changes in personal circumstances.

We can expect, however, that in some cases the reason for the decision to churn will leave an imprint in the data prior to the event. This could be in the form of certain patterns of complaints, or repairs, or other warning signs in the pattern of customer behaviour. In these cases, which we focus on in this paper, we may be able to model and therefore detect situations which will likely result in churn.

The remainder of the paper will be structured as follows. The next section will present the related work which will be followed by the results from an HMM method

using different length customer histories, as motivation for the non-sequential representation which will be presented in section 4. This section will also contain results using KNN for churn prediction. The final section will conclude.

2 Related Work

At the base of all churn prediction methods is the data used, and here already there are many options. Demographical data (i.e data about the customer) can be used to predict churn, however this may be unsuitable for a number of reasons [12]. Alternatives are call pattern changes and contractual information [12] or customer repair/complaint/provision data [7]. This latter type of data is that used in the current paper.

Neural networks, regression trees and linear regression are compared with regards to their churn predicting potential on repair/complaint/provision data in [7]. The regression tree was found to be most accurate overall achieving 82% correct predictions. However linear regression was the most successful in predicting non-churners whereas the neural network was better in predicting churners. Similar data in a sequential representation encompassing months of a customers historical data is used in a k nearest sequence method in [11] to predict churn, with an improvement found over standard classification techniques which use only the last month of data.

In [12], contractual and call pattern data are used together with a decision tree (C4.5, see [10]) based combination method. The combination method is used to combat the skewed nature of the data; as there are many more non-churn than churn examples, trees are trained on subsets of the training data each of which contains all the churn examples but different samples of the non-churn examples. This gives a number of more balanced training sets on which the trees are trained. The individual predictions are combined via weighted voting. The popular combination methods of bagging [2] and boosting [6] are tested on a mixture of customer and contractual data in [8].

This paper will focus on churn prediction from repair/complaint/provision data. As customers interact with the service provider, certain details of these interactions are logged, and from these we can build customer histories by constructing a sequence of time-ordered events for each unique customer. For the purposes of this paper, each event is described by 5 features. The precise details of the features cannot be given for reasons of confidentiality, but can be described in general terms. The (anonymised) dataset is available on request. One of these features is more naturally categorical; it denotes the event as one of four different types one of which is churn. These categories were expressed numerically for use in a Mixture of Gaussians Hidden Markov Model, or MGHMM (see section 3), the other features are naturally real-valued. One takes positive integer values from zero to a few hundred, two are positive real valued from zero to a few tens, and the final one is real valued with range \pm a few tens about zero.

Common sense suggests that more recent events should be given more weight when trying to predict future customer behaviour. This problem is quite common when dealing with prediction from sequential data; what is the relevance horizon of the data you have? In order to discover the timeframe over which it is best to take events when constructing a customer history, we constructed training sets in which only the most recent

N events are considered, for $N = 3 : 10$. When necessary, a subscript will denote the lengths of sequences allowed, so as an example TR_{any} or TR_N for $N = 3 : 10$.

It was found (see the next section) that models trained on the shortest histories performed best. This motivates the approach taken in section 4, as a short history can be expressed in a non-sequential representation quite easily. This could have applications in any domain where a short relevance horizon applies, especially in the services domain.

3 A Sequential HMM Approach

One class of method that has seen wide use and success on sequential data are Hidden Markov Models (HMMs) (see for example [5]). For a review of machine learning methods for sequential data see [4]. The simplest form of HMM assumes discrete outputs. For each event only certain discrete outputs can be produced, with the probability of each output depending only on the hidden state of the system. As the data we have consists of four continuous features and one categorical feature, it is more naturally represented in a continuous space so a more flexible model called Mixture of Gaussians HMM (or MGHMM) which allows for this is more useful. I will not describe the method further due to space constraints; the references above contain descriptions of the standard MGHMM we use in this paper.

HMM's will be generated from the customer data, trained iteratively via the usual EM (expectation maximisation) algorithm [1]. Separate models are trained on churn and non-churn sequences, denoted by M_c and M_n respectively, and classification is performed as follows. Given a trained model, the probability that it would generate a given test sequence can be calculated. The sequence can then be classified according to which model has the highest probability of generating it, taking into account the class priors.

These models are highly sensitive to the initialization of the model. One way of reducing this dependency on a specific initialization is to train a number of models using different initializations, and then combine their predictions. We have done this in a relatively simple, rank based manner. For a given individual pair of models M_c, M_n , after calculating for each sequence the probability of churn, the sequences are ranked in order of descending probability. For each sequence, then, we have the ranks $r = r_1, \dots, r_N$ where N is the number of models to be combined. We define a function to map this vector of values onto the real numbers, and rank them again according to this new value. We tried a variety of simple functions, and settled on an inverse square function $s = \sum_i \frac{1}{r_i^2}$ though performance is not too sensitive to the form of this function so long as it increases sufficiently quickly for small r_i .

We then take the top P sequences as our predictions. Here we have a trade-off to decide between. A larger P means we detect more of the actual churn events, but at a higher error rate. This trade-off is summed up in Fig. 1. For example, if we choose to take the top 0.4% as churn predictions (the percentage of sequences which are churn in the training set), we can expect a correct identification rate of just over 0.3. However if we choose to take the top 0.8% as predictions, we can expect to predict more churn events correctly (about 33% more) but at the lower recognition rate of 0.2. The experimental work will be covered in more detail in the next section.

This ability to specify trade-off easily is one advantage of a rank-based approach. Instead of choosing a fairly arbitrary threshold above which we will classify a sequence as in danger of churn, we can specify the level of trade-off we require and allow the data to set the threshold. We could then use this threshold for later classification of single sequences in for example an on-line scenario.

3.1 Results/Discussion

The data used in these experiments was constructed as described in section 2. There are 8080 customer history sequences from which to build the training data, but the final number of training/testing sequences will depend on the restrictions we place on their length. Sequences were split 60-40 into training and testing sets.

In Fig. 1, the performance measure is the fraction of churn predictions which are correct. The basic HMM architecture used was 12 hidden states, with 4 gaussians per state. Transitions depend only on the previous state. The HMM toolbox of Murphy [9] is used to build and train the HMM, and the default training parameters are used, with 12 training iterations. These values were chosen on the basis of preliminary tests. Varying these by a few either way has little effect, with the exception of reducing the number of gaussians below 4, which degrades performance quite markedly. A likely reason for this is that one feature (the event type) takes 4 discrete values, meaning at least 4 gaussians are needed to model the relative probabilities of these in general. Diagonal covariance matrices could have been used in order to reduce the number of parameters to be estimated, however this was not done as the data used is such that there is likely to be correlations between some features.

As can be seen in Fig. 1 the combination method improves performance quite significantly. This serves to illustrate that even quite simple combination methods can provide

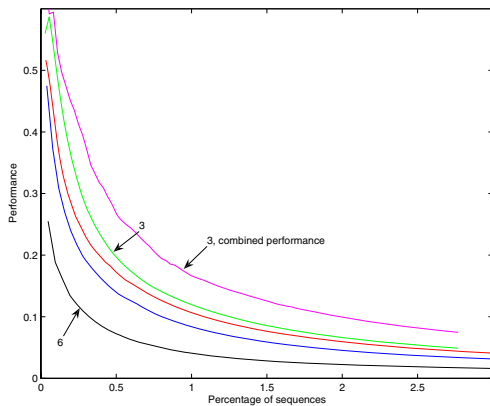


Fig. 1. The top line shows the combined performance using training sequences of length 3. Average performance of individual models plotted against percentage of sequences taken as predictions, for training sequences of length 3,4,5, and 6 are the plots (from top to bottom line).

a large benefit in real world applications. The length of sequence used in the histories can also be seen to have a large impact on performance, with shorter sequences of only the most recent historical events resulting in much better performance. This illustrates a point that it is still extremely important to choose the data correctly and represent it in the most suitable way. It is in this spirit that we will represent the data in a non-sequential manner in the next section.

In order to compare results from the HMM approach with those from the KNN method that follows, we will introduce a new performance measure. Performance will be measured using the following value:

$$G = \frac{p_{churn}}{p_{prior}} = \frac{c_{1|1}(c_{1|1} + c_{1|0} + c_{0|1} + c_{0|0})}{(c_{1|1} + c_{1|0})(c_{0|1} + c_{1|1})}$$

Where $c_{*|*}$ denotes the confusion matrix element, the first subscript being the predicted value (1 for churn, 0 for non-churn) and the second the actual. p_{churn} is the fraction of churn predictions which are correct, and p_{prior} is the prior probability of churn. This is used because, unlike the HMM, there is no natural way of specifying a tradeoff between number of predictions, and accuracy - the KNN method will simply give a set number of predictions. The metric is appropriate to the problem, as it is the ratio of the fraction of the methods churn predictions which are truly churn, to the fraction of examples that are churn. Thus if $G = 5$ say, this indicates that using the prediction method we make 5x more correct churn predictions than if we simply made predictions based on randomly predicting an observation to be churn in proportion to the prior probability of churn. It is the proportion of correct CHURN predictions, not the number of correct predictions absolute, which is important.

A further set of experiments was run for the shortest sequences. $Q = 2 : 20$ hidden states in the HMM were used for histories of length 2, for histories of length 3 this was only taken up to $Q = 10$. The dataset was again split 60-40 into training/testing sets and runs over 20 different splits were performed for each Q. The results are shown in Fig. 2. The number of predictions taken to give the g value is the same proportion of the testing set as are churn in the training set. It is also illustrative to look at the confusion matrices corresponding to some specific g-values indicated by arrows in [2](#).

$$g_1 = \begin{pmatrix} 27 & 46 \\ 85 & 21026 \end{pmatrix} \qquad g_2 = \begin{pmatrix} 13 & 64 \\ 24 & 21128 \end{pmatrix}$$

in the top right

$$g_3 = \begin{pmatrix} 29 & 43 \\ 43 & 21067 \end{pmatrix} \qquad g_4 = \begin{pmatrix} 22 & 54 \\ 54 & 21098 \end{pmatrix}$$

and the bottom two are

$$g_5 = \begin{pmatrix} 19 & 55 \\ 78 & 17700 \end{pmatrix} \text{ and } g_6 = \begin{pmatrix} 24 & 49 \\ 49 & 17903 \end{pmatrix}$$

It can be seen that though the g-value is actually increasing for larger Q and sequences of length 2, few churn predictions are actually being made at higher values (see g_1 compared to g_2) making the model less useful. The drop in g-value for higher Q for the

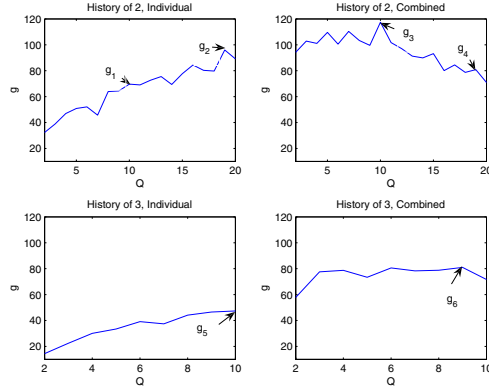


Fig. 2. G value vs. Q for individual and combined HMM predictions, using histories as labelled

combined method is probably related to this low number of churn predictions made by the individuals, and even though g is lower the combination method keeps the number of correct churn predictions made to a more useful level.

As can be seen again, the combination method clearly outperforms the individual models. The even shorter sequences of length 2 also outperform those of length 3, giving further motivation to our attempt at a non-sequential representation. This is the subject of the next section.

4 A Non-sequential KNN Approach

Instead of representing the customer history as a sequence, and making the implicit assumption that each event is related to the one before as in a HMM based approach, we may try to represent the data non-sequentially. In this case we make a slightly different assumption, which is that while the decision to churn is based on previous events, the previous events are not necessarily related to each other. Which is truer is debatable, it is easy to imagine scenarios where either could be the case. However there is no doubt that non-sequential data is easier to deal with. All the classical techniques such as KNN, parzen, tree, and support vector classifiers can be used, we chose KNN as an illustration as it performed better in preliminary tests, due to its suitability for problems when the classes are highly imbalanced. This suitability stems from the fact that by choosing K appropriately the number of data points contributing to the classification can be limited so that points of the more prevalent class do not always swamp the minority class. The fact that the churn examples tend to be a little more clustered than the non-churn also contributes to making KNN an appropriate choice.

A non-sequential dataset could be made from the above event histories by either averaging over the events in a sequence for each feature, to give the non-sequential feature values, or by creating new features to represent the features for different events. This latter would give Nk features for sequences of length N and events with k features.

It is the first method that we chose to use, though the second may be worth looking at in the future. The new features can be thought of as recording information answering questions like 'were things provided late during the last few events?', 'did the last few events take long?'. This is still highly useful information, what we lose is information on which event, for example, most of the delay/time was due to, or if it was spread over more than one. The second method of creating the new features would retain this information, but at the cost of creating many more features.

We show that when the relevance horizon for a sequential dataset is quite small, it is possible to get good results using classical techniques on a non-sequential representation. Guided by the results in the previous section which pointed to only the few most recent events being relevant, we chose to average features as little information is lost averaging a feature over just a few events.

Our features have some similarities with auto-regressive models. An AR model models a time series entry as being a linear combination of previous entries in the time series, possibly with a noise component:

$$X_t = c + \sum_{i=1}^{t-1} \alpha_i X_{t-i} + \epsilon_t$$

c is a constant and ϵ is a white noise component. We use a similar construction, but not in a predictive sense - we rather use it to construct a single feature which is a linear combination of feature entries in a time series:

$$X = c + \sum_{i=0}^{T-1} \alpha_i X_{T-i}$$

where T is the length of the series. Thus far we have used a very simple set of coefficients - the first τ α_i are $\frac{1}{\tau}$, the rest zero. An interesting extension would be to look at other sets of coefficients, perhaps exponentially decaying in timestep.

4.1 Method

The non-sequential dataset is constructed by averaging the features of the last τ events of the sequence, not including the last, label-defining event. The event type of this last event is used to label the data point as churn or non-churn. Only 3 features were included. These are event type (churn, complaint, repair and order are given the values 1,2,3 and 4 respectively), event duration (can be zero if not known or is not applicable), and promise (if something was promised, how early it was achieved; it is negative if that something was late. It can be zero if not relevant). These were chosen from a common sense view of what factors would be most likely to influence someone to churn, and from the results of preliminary experiments.

This dataset is split 60-40 into training and testing sets, and a simple K nearest neighbor algorithm is used to perform the classification. A nearest neighbor algorithm was chosen as it deals well with datasets such as this where the prior probabilities of the classes are highly imbalanced. The HMM is very computationally expensive to train, but it is cheap to calculate predictions when trained. In comparison, the KNN costs

nothing to train, however it can be very expensive to calculate very large numbers of predictions. There are many methods available in the literature to increase the efficiency of such nearest neighbor searches though, for just one example see [3]. Results, and some discussion and interpretation, follow.

4.2 Results/Discussion

The first three subplots in Fig. 3 show the results on datasets averaging the features of 1,2 and 3 events respectively, using 1-7 nearest neighbors for classification. The next event in the sequence is predicted. Performance is measured using the g value presented in section 3.1.

The final subplot shows the performance when trying to predict two events into the future, using a history of 2 and 1-9 nearest neighbors. This can be seen to be much less successful, showing that knowledge of the most recent event is very important. Using histories of other lengths to predict two events into the future also results in bad performance, and so is not shown.

From the above figures, it can be seen that an event history of 2 gives optimal performance using this method, and that taking simply the last event is totally inadequate for prediction. This shows that it is necessary to take into account the sequential nature of the data, even if only over a short time. An event history of 3 performs well, but worse than the shorter time period. This shows that the most relevant events in a customers history are the last two or three, as intuition would support. Also the non-sequential

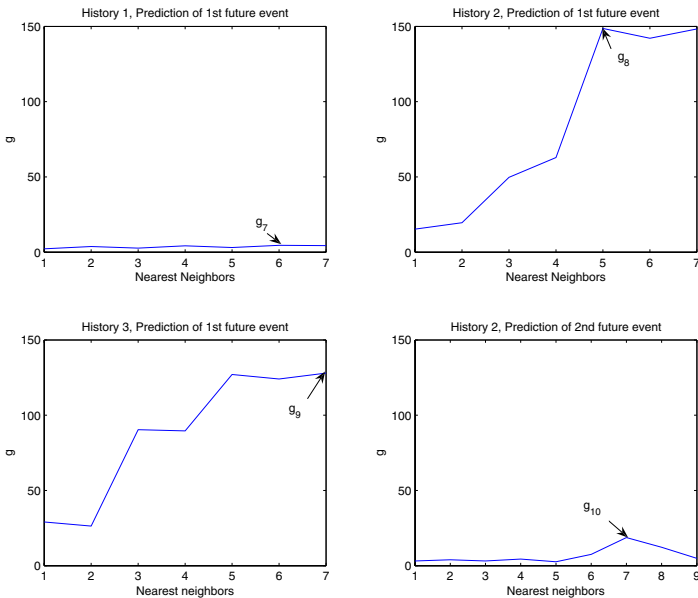


Fig. 3. Performance vs NN for histories and prediction time frame as labelled

representation is less suitable for longer sequences. Performance does not seem to be overly dependant on NN number, with a nearest neighbor count of over 5 performing well.

Again looking at the confusion matrices gives a little more insight. From top left to bottom right, the g values indicated are:

$$g_7 = \begin{pmatrix} 26 & 40 \\ 1278 & 13711 \end{pmatrix} \quad g_8 = \begin{pmatrix} 28 & 36 \\ 16 & 14866 \end{pmatrix}$$

$$g_9 = \begin{pmatrix} 9 & 57 \\ 7 & 14948 \end{pmatrix} \quad g_{10} = \begin{pmatrix} 3 & 51 \\ 34 & 12384 \end{pmatrix}$$

From g_7 we can see that although prediction from the last event detects churn quite well, there are very many false positives too resulting in a low g . From g_8 we see that including an extra event into the history has little effect on the number of correct churn predictions, but vastly reduces the number of false churn predictions, improving both specificity and sensitivity thus making this a much more useful tool for practical churn prediction. Looking at the confusion matrix g_3 for the HMM, we see that there is a decrease in false churn prediction compared to this too, while maintaining a very similar level of correct churn prediction. A third event in the history can be seen in g_9 to reduce churn predictions markedly, both correct and false. This lowers the sensitivity drastically, and in this case the number of churn predictions made is too low to be really useful, compared to using just 2 event histories.

Trying to predict more than one event into the future can be seen to result in very few churn predictions.

We can attempt to interpret what these results could mean in real terms by looking more closely at the nature of the data we have. Roughly half of all the churn examples correspond to sequences in which the last two events are complaints, which is revealing in itself, although it shouldn't really be surprising. Almost all the churn examples correspond to event sequences in which the last two events have been of the same type, and many of them where the last two events are quite similar. These observations could be interpreted as indicating that a customer does not like to have to do the same thing twice when dealing with the service provider, especially when that thing is a complaint. Churn examples are also quite closely clustered, indicating that complaints falling into a few distinct, well defined subclasses may be especially likely to provoke a churn response.

5 Conclusions

We have proposed, based on observations from a HMM method, that only the most recent events in a customers history have an effect on the future behaviour of that customer, and shown that a short sequence of events corresponding to a recent history can be represented easily in a non-sequential way. This allows the use of all the tools available for simple, non-sequential pattern recognition, and we show that a K-nearest neighbor algorithm performs well on this data. This provides much better performance and potentially reduced computational complexity over the HMM methods. We explain the success of this method by noting that many churn events when represented in this

way lie in a few small, dense clusters, and observe that many churn events follow a history of two events of the same type, often with similar feature values. This indicates that perhaps having to do the same thing twice, especially with regards to a complaint, often leads to churn.

References

1. Bilmes, J.: A gentle tutorial on the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models (1997)
2. Breiman, L.: Bagging predictors. *Machine Learning* 24(2), 123–140 (1996)
3. Chen, Y.-S., Hung, Y.-P., Yen, T.-F., Fuh, C.-S.: Fast and versatile algorithm for nearest neighbor search based on a lower bound tree. *Pattern Recogn.* 40(2), 360–375 (2007)
4. Dietterich, T.G.: Machine learning for sequential data: A review. In: Caelli, T.M., Amin, A., Duin, R.P.W., Kamel, M.S., de Ridder, D. (eds.) *SPR 2002 and SSPR 2002*. LNCS, vol. 2396, pp. 15–30. Springer, Heidelberg (2002)
5. Duda, R., Hart, P., Stork, D.: *Pattern Classification*. John Wiley and Sons, Chichester (2001)
6. Freund, Y., Schapire, R.E.: Experiments with a new boosting algorithm. In: *Proceedings of the 13th International Conference on Machine Learning*, pp. 148–156. Morgan Kaufmann, San Francisco (1996)
7. Haddon, J., Tiwari, A., Roy, R., Ruta, D.: *Churn prediction: Does technology matter* (2006)
8. Lemmens, A., Croux, C.: Bagging and boosting classification trees to predict churn. *Journal of Marketing Research XLIII*, 276–286 (2006)
9. Murphy, K.: A hmm toolbox for matlab, <http://www.cs.ubc.ca/~murphyk/software/hmm/hmm.html>
10. Quinlan, J.R.: *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc., San Francisco (1993)
11. Ruta, D., Nauck, D., Azvine, B.: K nearest sequence method and its application to churn prediction. In: Corchado, E., Yin, H., Botti, V., Fyfe, C. (eds.) *IDEAL 2006*. LNCS, vol. 4224, pp. 207–215. Springer, Heidelberg (2006)
12. Wei, C.-P., Chiu, I.-T.: Turning telecommunications call details to churn prediction: a data mining approach. *Expert Systems with Applications* 23, 103–112 (2002)
13. Yan, L., Miller, D.J., Mozer, M.C., Wolniewicz, R.: Improving prediction of customer behaviour in non-stationary environments. In: *Proc. of Int. Joint Conf. on Neural Networks*, pp. 2258–2263 (2001)

Dialectics-Based Knowledge Acquisition – A Case Study

Cecilia Zanni-Merk and Philippe Bouché

LGECO – INSA de Strasbourg – 24 boulevard de la Victoire – 67084 Strasbourg – France
{cecilia.zanni-merk, philippe.bouche}@insa-strasbourg.fr

Abstract. This article presents our proposition of a methodology for knowledge acquisition based on dialectics. In fact, the central concept in dialectics is a “contradiction”, which declined according to the inventive design principles, can be considered as a set of Elements, Parameters and Values - values that need to show the opposite aspects of the contradictions. Using these approaches for knowledge acquisition permitted us to obtain very satisfying results for solving a problem of software analysis.

Keywords: knowledge acquisition, knowledge management, dialectics, inventive design.

1 Introduction

Knowledge acquisition includes the elicitation, collection, analysis, modelling and validation of knowledge for knowledge engineering and knowledge management projects.

One of the most important issues in knowledge acquisition is that knowledge is in the heads of experts and that they have vast amounts of it. These experts manipulate tacit knowledge, without being able to articulate how they do so. Another constraint is that tacit knowledge is hard (impossible) to describe [1].

Because of these issues, techniques are required which focus on essential knowledge and that can capture tacit knowledge, that allow knowledge to be validated and maintained and that allow non-experts to understand it [1]. Examples of these techniques include the use of ontologies and of problem-solving models.

Methodologies have been introduced that provide frameworks to help guide knowledge acquisition activities and ensure that the development of each expert system is performed in an efficient manner. A leading methodology is CommonKADS [2]. At the project management level, CommonKADS advises the use of six high-level models. To aid in the development of these models, a number of generic models of problem-solving activities (a number of generic problem-solving activity models) are included. Each of these generic models describes the roles that knowledge plays in tasks, hence providing guidance on what types of knowledge to focus upon. As a project proceeds, CommonKADS follows a spiral approach to system development, providing for rapid prototyping of the system, so that there is more flexibility in dealing with uncertainty and change.

A second important development is the creation and use of ontologies [3]. Although there is a lack of unanimity in the exact definition of the term ontology, it is generally

regarded as a formalised representation of the knowledge in a domain taken from a particular perspective or conceptualisation. The main use of an ontology is to share and communicate knowledge, both between people and between computer systems. A number of generic ontologies have been constructed, each having an application across a number of domains which enables the re-use of knowledge. In this way, a project does not need to start with a blank sheet of paper, but with a number of skeletal frameworks that can act as predefined structures for the knowledge being acquired. As with the problem-solving models of CommonKADS, ontologies also provide guidance to the knowledge engineer in the types of knowledge to be investigated.

These approaches present some disadvantages in their use, before really beginning the elicitation process. On the one hand, if an ontology-based approach is used, a model of the whole domain (the base ontology) is needed, before beginning the “instantiation” of it. On the other hand, while using problem solving models, even if there is no need to know the base ontology in advance (although it is strongly recommended), the elicitation runs the risk of quickly becoming “chaotic”. Confronted with these issues, we decided to try the use of dialectics-based techniques to guide the experts during the knowledge acquisition process, with the goal of counteracting the difficulties we have described above.

In this article, therefore, we will present our approach to knowledge acquisition, based on inventive design techniques, which are based on dialectical thinking. Section 2 briefly describes dialectics and the inventive design concepts that will be used. Section 3 describes the case study conducted with this new knowledge acquisition technique and Section 4 presents our conclusions.

2 Dialectics and Inventive Design

Dialectics is a philosophical school with roots in the old Greek philosophy, represented by e.g. Heraclitus. The philosophy grows out of the Hegelian discussion about the relation (or contradiction) between ideas and reality.

The key concept of dialectics is contradiction. A contradiction consists of two aspects, which are mutually dependent and opposed to each other at the same time. All complex phenomena consist of several contradictions, one of them dominating the others and characterizing the phenomenon. This one is the principal contradiction. Furthermore, all processes consist of a movement of contradictions from the beginning to the end. Through time, the principal contradiction may change. According to dialectics, the causes of change and evolution are: (1) the changing relation between two aspects of each contradiction and (2) the changing relation among the contradictions of a certain phenomenon [4].

Regarding Inventive Design, we are interested in TRIZ [5, 6] (the Russian acronym for Theory of Inventive Problem Solving). Its core approach is to structure inventive thinking. It is based on studies of several hundreds of thousands of patents and it has outlined typical general problems and their general solutions.

TRIZ is primarily about technical and physical problems, but is now being used on almost any problem or situation. The key to success in TRIZ is the fact that (technical) systems evolve in similar ways, and by reducing any situation and problem to a functional level, we can apply almost standard solutions and problem solving techniques, even from dissimilar industries.

The complete list of TRIZ components is very long indeed, and includes laws, methods, and various lists of principles. We will focus, here, on Contradictions, one of the essential TRIZ axioms. It stipulates that every inventive problem, in its formulation, may be reduced to a contradiction. Contradictions have, individually, three types of components: *Elements*, *Parameters* and *Values*.

Elements are the constituents of systems. In written texts, they are usually expressed by nominal groups or object complements (for example, in the sentence “the hammer hits the nail”, “hammer” is the element).

Parameters qualify the elements by assigning a certain specificity to them that, associated with the element, translates explicit knowledge of the observed domain. They are often expressed by nouns, object complements or adverbs. Their expression is multiple, and sometimes contradictory when established by different experts. They are classified in two categories:

- *Control Parameters*, the designers have the possibility of modifying them (the designer may choose to have a hammer with a big or a small volume, in this case, “volume” is a control parameter).
- *Evaluation Parameters*, whose nature resides in their capacity to evaluate the positive aspect of a choice made by the designer. Designing a hammer with a big mass assures that hitting the nail will be easier. In this way, “ease of hitting” becomes an evaluation parameter.

Value is mainly the adjectives used to qualify a parameter (in the example “the volume of the hammer has to be big”, “big” is a value).

The fundamental aspect of contradictions resides in the opposition of the values and in the fact that we have to make the two opposite values explicit. If in a certain state, a value *V* implies positive aspects, then it is indispensable to investigate the opposite of *V*, to make the contradictory aspects of the analysis evident. For example, a hammer with a big volume implies “ease of hitting” and a hammer with a small volume implies “ease of handling”.

Several techniques are used for making contradictions and its components readily evident while formulating the problem to be solved. We have used the so-called “Part 0 of ARIZ¹” [6] and the “Problem Flow Network” (PFN) approach [7], a component of OTSM-TRIZ [8].

2.1 Part 0 of ARIZ

The first stages of the creative inventive process are devoted to choosing the problem and *redefining its conditions*. Most of the time, the original statement of the problem is imprecise, and occasionally even incorrect. For example, for the statement “we need to find a method to provide *such* and *such* a function”, it might be better to eliminate the necessity for this function all together. Very often, the bypass concept is more productive than the direct one. Part 0 of ARIZ establishes the final goal of the solution, investigates the possible use of a bypass solution and redefines the conditions for them (both the direct and the bypass solution). Afterwards, the requirements are deliberately increased [6].

¹ ARIZ is the Russian acronym for the “Algorithm of Inventive Problem Solving”.

Another important point is the definition of “performance”, the idea is to determine the best ratio “technical / economical characteristics” that can be achieved once the solution is implemented.

Other points include the analysis of patents related to the problem and the use of the STC Operator. Psychological inertia is caused not only by the terminology describing the object, but also by the customary space/time imaging of the object: its *size*, as well as the *duration of its action*. There is another measurement of an object’s mental image, *cost*. The STC operator is a sequence of mental experiments helping to overcome these conventional images of an object, by considering the successive changes in the problem when changes are made in three parameters: size (S), time (T) and cost (C) [6].

2.2 The Problem Flow Network Approach

Complex problem situations are characterized by a large set of parameters and interference of the elements involved in the problem. Using the concept of the network PFN approach allows the representation and analysis of several contradictions and problems at once. This approach is based on the construction of several networks, and in particular, the network of Problems and Partial Solutions.

A usual starting point for the construction of this network is a list of problems, difficulties, inconveniences and questions relevant to the situation. This list is made by professionals dealing with the problem. This network is an oriented graph, whose nodes represent the problems while the edges represent the links between the problems. The graph also contains nodes that present known Partial Solutions of certain problem nodes. Problems that are generated by Partial Solutions are also presented. Problems are linked to the evaluation parameters of the system. Partial solutions are linked to the control parameters the designers may act on.

The construction of the problem network is an iterative process. Sometimes, just the building of it permits a final conceptual solution, which can be accepted and implemented, to be found.

Once the problem network becomes stable enough and the key problems are extracted (this extraction may be made by the analysis of the properties of the graph that represents the network [9]), the set of contradictions that are behind the set of key problems can be disclosed.

3 Case Study: Analysing the Productivity of a Manufacturing Line

The goal of a project with an industrial partner was the development of a software tool for detection and analysis of the causes of poor productivity in a manufacturing line. The detection is done by the processing of timed data coming from a data acquisition system [10].

After calculation with these data, we are able to detect phenomena that may induce a lack of productivity on the manufacturing line.

A phenomenon is the expression of a particular behaviour of the production line which has duration. Examples of phenomena include, lack of components, stock saturation or slow handling speed.

The set of causes for the lack of productivity in a manufacturing line is expressed as a behavioural model. A behavioural model is a set of sequences of correlations, where a correlation is expressed as an ordered couple of phenomena with time constraints (Figure 1).

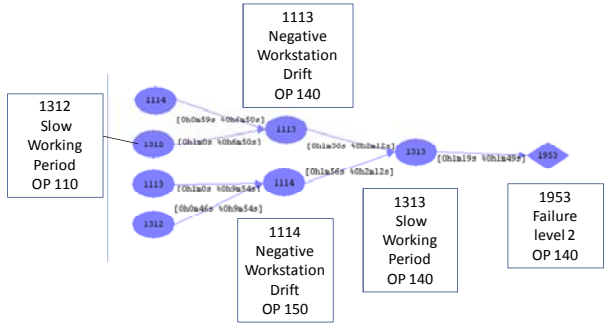


Fig. 1. Example of a behavioural model

Data coming from the data acquisition system are processed by three routines:

- *Routines for pre-processing*, which aim at filtering (“cleaning”) raw data coming from the data acquisition system.
- *Routines for phenomena detection*, which are a set of algorithms that take data as inputs and aim at computing occurrences of phenomena. To do this work, each algorithm uses a set of arguments and thresholds specified by experts.
- *Routines for phenomena correlation*, which are a set of algorithms that take a log of phenomena as an input and aim at identifying the behavioural models. These routines perform probabilistic computations for identifying correlations among phenomena and calculating their associate time constraints. There are a set of intermediate outputs whose specific characteristics can be used to evaluate the quality of the results.

These three routines are executed in sequence. Once their intermediate and final results are computed, the production expert and the data-mining expert can evaluate them. If the quality of these results is not satisfying, the experts may decide to change certain inputs to the phenomena detection and phenomena correlation routines and re-launch all or part of the process.

The main drawback here is that experts are used to evaluating the quality of the results only in a qualitative way, such as “it is good” while not being able to clearly say what the meaning of “good” is.

During the setup period of the software for a new production line, this “trial and error” process is a waste of the experts’ time, as they test different configurations until they get satisfying results. The search for the best configuration of arguments has a linear complexity, or in other terms, it is $\mathcal{O}(n)$.

3.1 The Knowledge Acquisition Process

We have decided to use Part 0 of ARIZ and the Network of Problems approach as a basis for the interviews with the data-mining and the production experts, in order to find the right input arguments for routines having optimum quality behavioural models.

Several iterations were needed and each of them on every approach gave more insight into the description/use of the other one.

As the object of our study was software routines, it was quite a natural choice to consider input arguments as control parameters and outputs as evaluation parameters.

A special effort was made to define our measure of performance², as

$$P = \frac{\text{structure of the MPSR tree} \rightarrow \text{optimum} \ \&\& \ \text{cover ratio} \ (\%) \rightarrow 100\%}{\text{number of adjustment executions} \rightarrow 0}$$

and to associate a unique evaluation parameter to a problem node in the network, and a unique control parameter to a partial solution node.

Figure 2 presents an extract of the problem network we have built, where interactions among problems and partial solutions are stated.

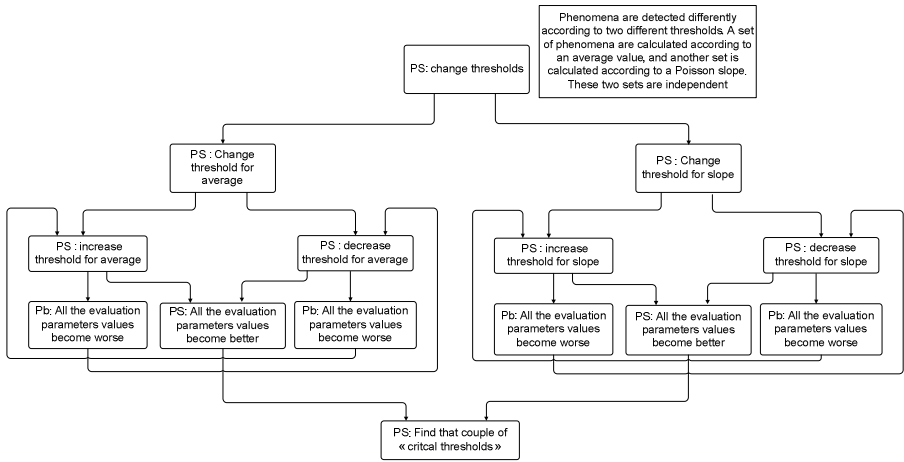


Fig. 2. A subset of the network of problems. Boxes whose prefix is Pb represent problems; boxes whose prefix is PS represent partial solutions.

These iterations made the experts aware of links among control and evaluation parameters that were not evident until then. In particular, the “satisfying” or “unsatisfying” intervals associated with the evaluation parameters.

This awareness permitted the definition of new heuristics for searching the good input arguments for the routines. This new strategy’s complexity is $\mathcal{O}(\log n)$. More than that, this new heuristic search does not need the re-launching of the phenomena correlation routines, once the good inputs of the phenomena detection routines have been found. Figure 3 shows the comparison of the two methodologies used by the experts to set up the software for a new production line, the old and new heuristics.

² See [11] for a definition of the evaluation parameters in the numerator of *P*.



Fig. 3. Comparison of the two heuristics. Notice that several feedback links have disappeared.

4 Conclusions

This article has presented a new approach to knowledge acquisition, based on inventive design principles.

As it was said in the Introduction, experts have their own cognitive models and make operations intuitively. That is tacit knowledge that they are not able to formalize.

The utilization of these inventive design principles and especially, of Part 0 of ARIZ, has forced the experts to:

- Express quantitative criteria and not qualitative ones for evaluation of the results.
- Simply explain the process by using conventional representation models such as input/output models or cause-effect relationships.

The construction of the network problem was an important phase to identify the quantitative parameters that can help make decisions. After several iterations, its elicitation

has permitted the obtainment of the list of evaluation and control parameters. The main issues were the identification of:

- The factors involved in data analysis and therefore in the construction of the behavioural models.
- The criteria for evaluating the quality of the data analysis.

The results we have obtained are very satisfying. The new methodology can be applied to any manufacturing line to be analysed by our software with the goal of understanding the causes of poor productivity. We have been able to extract tacit knowledge from the experts.

This experience has showed the interest of using inventive design approaches to formalise expert knowledge. Our future applications include the use of this technique for knowledge acquisition on territorial development of harbours.

References

1. Milton, N.R.: *Knowledge Acquisition In Practice: A Step-By-Step Guide*. Springer, London (2007)
2. Schreiber, G., Hakkermans, H., Anjewierden, A., de Hoog, R., Shadbolt, N., Van de Velde, W., Wielinga, B.: *Knowledge engineering and management - The CommonKADS methodology*. MIT Press, Cambridge (2000)
3. Guarino, N.: The Ontological Level. In: Casati, R., Smith, B., White, G. (eds.) *Philosophy and the Cognitive Science*, pp. 443–456. Holder-Pivhler-Tempsky, Vienna (1994)
4. Cassin, B. (ed.): *Vocabulaire européen des philosophies*, p. 306. Le Robert & Seuil, Paris (2004)
5. Altshuller, G.: *Creativity as an Exact Science*. Gordon and Breach Scientific Publishers, New York (1984)
6. Altshuller, G.: *TRIZ The innovation algorithm; systematic innovation and technical creativity*. Technical Innovation Center Inc., Worcester, MA (1999)
7. Khomenko, N., De Guio, R., Lelait, L., Kaikov, I.: A Framework for OTSM-TRIZ Based Computer Support to be used in Complex Problem Management. *International Journal Computer Applications in Technology* 30(1/2), 88–104 (2007)
8. Khomenko, N., Kucharavy, D.: OTSM-TRIZ problem solving process: solutions and their classification. In: *TRIZ Future Conference*, Strasbourg, France (2002)
9. Cavallucci, D., Rousselot, F., Zanni, C.: Assisting R&D activities definition through problem mapping. *CIRP Journal of Manufacturing Science and Technology* 1(3), 131–136 (2009)
10. Bouché, P., Zanni, C.: Pro@ctiF: an expert system to improve performance of production lines. In: *2008 Summer Computer Conference*, Edinburgh, Scotland (2008)
11. Bouché, P., Le Goc, M., Coinu, J.: A Global Model of sequence of discrete event class occurrences. In: *ICEIS 2008, 10th International Conference on Enterprise Information Systems*, Barcelona, Spain (2008)

Automatic Extraction of Hyponymy-Hypernymy Lexical Relations between Nouns from a Spanish Dictionary

Rodolfo A. Pazos R., José A. Martínez F., Juan J. González B.,
María Lucila Morales-Rodríguez, and Jessica C. Rojas P.

Instituto Tecnológico de Ciudad Madero
Av. 1º. de Mayo esq. Sor Juana Inés de la Cruz, 89440, Cd. Madero, Mexico
r_pazos_r@yahoo.com.mx, {jose.mtz,lmoralesrdz}@gmail.com,
{jjgonzalezbarbosa,isc_jessicarojas}@hotmail.com

Abstract. In this paper a method is presented which permits to automatically extract lexical-semantic relations between nouns (specifically for concrete nouns since they have a well structured taxonomy). From the definitions of the entries in a Spanish dictionary, the hypernym of an entry is extracted from the entry definition according to the basic assumption that the first noun in the definition is the entry hypernym. After obtaining the hypernym for each entry, multilayered hyponymy-hyperonymy relations are generated from a noun, which is considered the root of the domain. The domains for which this approach was tested were zoology and botany. Five levels of hyponymy-hypernymy relations were generated for each domain. For the zoology domain a total of 1,326 relations was obtained with an average percentage of correctly generated relations (precision) of 84.31% for the five levels. 91.32% of all the relations of this domain were obtained in the first three levels, and for each of these levels the precision exceeds 96%. For the botany domain a total of 1,199 relations was obtained, with an average precision of 71.31% for the five levels. 90.76% of all the relations of this domain were obtained in the first level, and for this level the precision exceeds 99%.

1 Introduction

Language in written form has become a valuable means of manifestation of natural language, which is being used as vehicle for communicating acquired knowledge, as well as ideas and feelings. Since the advent of computers, texts in electronic format have become one of the main forms for information exchange.

Natural language processing (NLP), in its attempt for automatically manipulating electronic texts, performs tasks such as: filtering, classification, and information retrieval and extraction. Such tasks use lexical resources as an aid for improving the performance of automatic text processing. Some of the lexical resources used are ontologies, which are databases that store concepts and relations for representing, organizing and understanding a knowledge domain, and have as their main objective to explicitly establish one or several relations among words of a language. Due to their valuable contribution, they have attracted the interest for studying and proposing methods that permit building such resources, either manually or automatically. The

use of an ontology seems to be an adequate form for incorporating lexical-semantic knowledge into an NLP system [1].

Ontologies originally were generated manually [2]; however, manual acquisition of knowledge for these resources has proven highly costly in terms of human time and effort [3]. Besides, it would be almost impossible or very complicated to generate in such a way a resource that incorporated all the lexical knowledge of a language. For example, for the construction of the famous lexical database for English, WordNet, countless hours have been dedicated and still not all the lexical knowledge for that language has been incorporated [4].

As a result, several investigations have been carried out for automating such process. Such works had less ambitious objectives; i.e., they focused on one or several relations and not all of them. Additionally, they were limited to dealing with specific domains. Some of the first investigations dealt with very structured text types, such as English and Spanish dictionaries, among other languages.

Relations among words in an ontology are known as lexical-semantic relations. These relate words according to their meaning [5]. Among the lexical-semantic relations on which research work has been carried out, stand out synonymy, antonymy, meronymy-holonymy, and hyponymy-hypernymy.

This work focuses on extracting hyponymy-hypernymy relations between nouns which are entries of a Spanish dictionary. A hyponym is a word whose meaning is included in that of another whose meaning is more general, which is called hypernym. For example, cat is a hyponym of felis.

2 Related Work

The research works that have been carried out to automatically extract hyponymy-hypernymy relations have used several techniques. One of these relies on the notion that there exist key fragments of the text (words, punctuation marks or both) that indicate the presence of these relations in a document. These text fragments may appear repeatedly, so that they can be generalized into patterns (lexical-semantic patterns). When applying these patterns to text collections, instances of the hyponymy-hypernymy relation can be extracted from text documents. These patterns can be manually constructed or automatically generated, as proposed in [6].

In [7] an algorithm for extracting semantic relations is presented. This algorithm takes as input a set of seeds of a semantic relation for extracting instances that belong to such relation. The iterative mechanism employed permits it to learn new lexical-syntactic patterns at each iteration; and consequently, the set of instances that it can recognize grows. In this work, the confidence of instances and patterns was evaluated considering the mutual information that exists among patterns and relation instances.

The investigation presented in [5] applies the technique proposed in [7], which has the following drawbacks: the patterns identified from unstructured texts from the web are exclusively lexical, which makes difficult to detect the lexical components in a sentence (such as articles, verbs, etc.), and additionally, due to the nature of the lexical information, a large number of false hyponyms can be obtained after applying the lexical patterns to a text collection.

The advantage of the techniques based on patterns is that they are very reliable. For example, in [7] a precision of 85% is reported on a random sample of 20 instances out of 200. However, the main disadvantage of these methods is that they need a very

large corpus in order to find a large enough number of patterns that describe all the possible hyponymy relations that may exist [8].

Another technique assumes that dictionaries in a format readable for a computer store explicit knowledge in structured form, which can be extracted by retrieving instances of semantic relations, including hyponymy relations. This technique relies on the following notion: the hypernym of a dictionary entry is expected to be found in the first noun phrase of its definition.

Due to the generally regular structure of dictionaries, the results obtained are usually high. For example, in [3] a precision of 87% is mentioned, 77% is reported in [9], and a precision of over 90% is claimed in [10]. Therefore, it can be concluded that this method has better precision. However, this technique does not consider specific terms for a domain. This happens because dictionaries are almost always very general resources that deal with terms usually utilized in many different domains. This inconvenience generated interest in exploring other approaches for extracting hyponyms.

Another of the techniques proposed for extracting hyponymy-hypernymy relations consists of identifying the grammatical role of each word that constitutes a sentence. In [2] this method was tested on documents from the zoology domain for Spanish. A general document and a specialized one were used for contrasting the relations found in each of the documents, and thus assessing the effectiveness of this approach. One of the problems found was that obvious relations were not identified, such as *dog is an animal*; thus, the possibility of finding the missing relations in a dictionary was proposed.

Another technique proposes the use of clustering based on distributional similarity between terms for taxonomy extraction [11]. The authors claim that the accuracy of the resulting taxonomy improves considerably when building the taxonomy using several different languages (English, German, French and Spanish) over a monolingual approach.

3 General Description of the Approach

The basic assumption of the proposed approach is that the hypernym of a noun defined in a dictionary is the first noun found in its definition. Fig. 1 shows the general structure of the proposed approach for identifying hyponymy-hypernymy relations between nouns.

As previously mentioned, a dictionary is a particularly adequate source for constructing lexical resources. In order to avoid keying in information from a printed dictionary, it was decided to use an electronic version: LEXI-K [12], which is the Spanish dictionary used for this work.

The process starts by identifying the entries (words defined in a dictionary) that are nouns, from information stored in the electronic dictionary. Afterwards, the hypernym for each noun identified is extracted from its definition. In order to accomplish this, it is necessary to obtain first the syntactic category (noun, verb, adjective, etc.) of each of the words involved in its definition.

Afterwards, the root noun for some domain of interest is determined. To this end, the hypernyms of some "seed" nouns that belong to the domain are considered; then the hypernyms of these hypernyms are determined recursively, until a noun is found with no hypernym, which constitutes the root. Finally, the hyponymy-hypernymy relations among all the nouns for a particular domain are obtained starting from the root noun for the domain, similarly to a taxonomy structure.

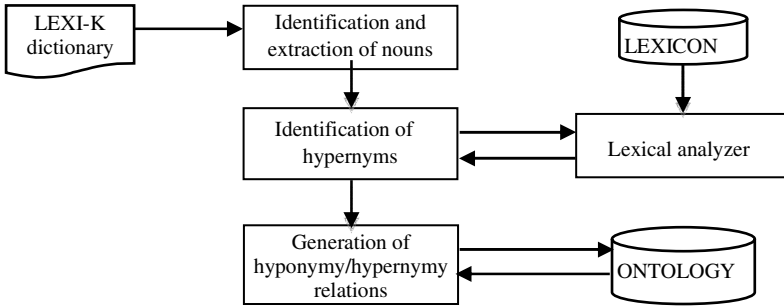


Fig. 1. General structure of the approach

3.1 Noun Identification

For identifying the entries that are nouns it was first necessary to find out how LEXI-K stores its entries. Consequently, the LEXI-K interface was studied and those files that interact with it were identified.

Since the files that store the LEXI-K information are coded in binary format, a program was implemented for identifying and extracting the entries that are nouns, and storing them in a file called Nouns.

3.2 Hypernym Identification

For identifying the hypernym in an entry definition, it is necessary to obtain automatically the structure of its definition. This was carried out using a lexical analyzer, which in turn uses a lexicon [13].

The process starts obtaining the first instance of file Nouns, from which the entry and its definition(s) is(are) obtained. It is important to mention that several nouns were tagged as invalid nouns (shown in Table 1), since they indicate the hypernym relation and, therefore, do not constitute hypernyms themselves.

After reading the first instance, the structure of its definition(s) is(are) generated. Next the first valid noun is looked for (excluding invalid ones) in each of its definitions, which according to our assumption is the entry hypernym. The hypernym found

Table 1. Invalid nouns considered for this approach

Invalid nouns	
<i>Especie</i> (species)	<i>Género</i> (genus)
<i>Suborden</i> (suborder)	<i>Subreino</i> (subkingdom)
<i>Grupo</i> (group)	<i>Individuo</i> (individual)
<i>Clasificación</i> (classification)	<i>Nombre</i> (name)
<i>Clase</i> (class)	<i>Variedad</i> (variety)
<i>Tipo</i> (type)	<i>Taxón</i> (taxon)
<i>Familia</i> (family)	<i>División</i> (division)
<i>Serie</i> (series)	<i>Orden</i> (order)
<i>Subtipo</i> (subtype)	<i>Colonia</i> (colony)

Table 2. Sample of tuples of the table of hyponym-hypernym pairs

Hyponym	Hypernym
<i>gato</i> (cat)	<i>mamífero</i> (mammal)
<i>león</i> (lion)	<i>mamífero</i> (mammal)
<i>margarita</i> (daisy)	<i>planta</i> (plant)
<i>roble</i> (oak)	<i>árbol</i> (tree)
<i>mesa</i> (table)	<i>mueble</i> (furniture)
<i>silla</i> (chair)	<i>asiento</i> (seat)
<i>trompo</i> (trochus)	<i>molusco</i> (mollusc)
<i>trompo</i> (top)	<i>juguete</i> (toy)

is transformed to its singular form if necessary, and it is assigned to the entry of the corresponding instance as its hypernym. Finally, both elements (hyponym and hypernym) are stored in a table (as shown in Table 2). This process is repeated for each of the instances in the file.

3.3 Extraction of Hyponymy–Hypernymy Relations

After identifying the hyponym-hypernym pairs (Table 2), relations among these pairs are identified, according to the strategy described next:

1. Choose a domain or particular area.
2. Identify the root noun of the domain.

The system is feed one or several nouns that belong to the domain, for which its(their) hypernym(s) is(are) determined from the hyponym-hypernym pairs; then the hypernym(s) of this(these) noun(s) is(are) determined. This process is repeated until a single noun is found for which no hypernym can be found. This last noun constitutes the domain root.

3. Look for hyponym-hypernym pairs that contain the root noun of the domain in its *hypernym column*.

If such hyponym-hypernym pairs were found, store them.

Select the element in the *hyponym column* of each of these pairs and tag it as *hyponym i*, where $1 \leq i \leq n$; i.e., *i* will acquire a value in the interval from 1 to *n* (the overall number of hyponym-hypernym pairs that contain the root noun of the domain in its *hypernym column*).

For each *hyponym i*

If no hyponyms have been obtained for *hyponym i*

Obtain *hyponyms k* for *hyponym i* (where $1 \leq k \leq m$; i.e., *k* will adopt a value in the interval from 1 to *m*, which indicates the overall number of hyponym-hypernym pairs that contain *hyponym i* in its *hypernym column*), similar to what occurred for the root noun of the domain. This process is applied to the rest of descendants.

4 Results

From the 43,379 nouns identified from the LEXI-K dictionary, the overall number of hyponym-hypernym pairs obtained was 52,819, from which a small sample of 190

pairs of the zoology domain and 190 pairs of the botany domain were selected for verifying their correctness. The percentage of correctly generated pairs (success rate) for the zoology domain was 97% and for the botany domain was 94%.

Incorrectly generated hyponym-hypernym pairs were studied in order to discover and eliminate the causes of imperfection from our method. The improved method was applied again to the LEXI-K dictionary, and 43,289 hyponym-hypernym pairs for nouns were obtained.

Table 3. Lexical relations identified for the zoology domain

Level	Relations obtained	Incorrect relations	Success rate
1	180	5	97.22%
2	511	8	98.43%
3	355	12	96.62%
4	142	80	42.33%
5	138	103	25.36%
Total	1,326	208	84.31%

With the hyponym-hypernym pairs, it was possible to identify lexical relations for the zoology domain as shown in Table 3 and for the botany domain, shown in Table 4.

Table 4. Lexical relations identified for the botany domain

Level	Relations obtained	Incorrect relations	Success rate
1	779	3	99.61%
2	147	85	42.17%
3	116	113	22.41%
4	90	80	15.55%
5	67	63	5.97%
Total	1,199	344	71.31%

The average success rate for the zoology domain was 84.31%, where the maximal value of 97.22% was obtained for the first level and the minimal value of 25.36% for the fifth level. The average success rate for the botany domain was 71.31%, with a maximal value of 99.61% for the first level and a minimal value of 5.97% for the fifth level.

The difference in the success rates obtained for the first and fifth domains can be explained by the fact that incorrectly generated relations obtained at one level induce more mistakes for the following levels.

It is important to point out that the main source of errors when generating relations among hyponym-hypernym pairs was polysemy, specifically when a word has an alternate meaning related to another word that belongs to a different domain. For example, for the entry *trompo* (*trochus*) shown in Fig. 2, the hypernym identified is *molusco* (*mollusc*), which was correctly obtained.

Entry	Grammatical Features		Definition
	Gender	Number	
<i>trompo</i> (trochus)	M	S	Marine gastropod mollusc , with conical tentacles on the head and conical shell, thick.

Fig. 2. Definition of entry *trompo* (trochus)

Entry	Grammatical Features		Definition
	Gender	Number	
<i>rezumbador</i> (humming top)	M	S	Kind of top that hums when spinning.
<i>trompa</i> (no English equivalent)	F	S	Large top that holds inside other small tops, which get off impetuously when thrown for spinning, all spin simultaneously.

Fig. 3. Definitions of entries *rezumbador* (humming top) and *trompa*

Unfortunately, in Spanish *trompo* has another meaning besides trochus, which corresponds to a toy top. Consequently, the hyponyms generated for *trompo* are *rezumbador* (humming top) and *trompa* (no English equivalent), which are incorrect for the zoology domain, as can be seen from the definitions shown in Fig. 3.

5 Conclusions

In the research work described in this paper, we propose extracting hyponymy-hypernymy relations for specific domains from a Spanish dictionary. This method was tested for the zoology and botany domains.

The basic assumption of this approach was shown effective; i.e., it is possible to use a simple general pattern, which allows to identify hypernyms of noun entries from their definitions. A software based on the approach proposed was implemented, which permits to automatically identify nouns hypernyms.

A total of 1,326 hyponym-hypernym pairs were obtained for the zoology domain and 1,199 for the botany domain. For the zoology domain, the percentage of correctly generated relation pairs was 84.31%. With respect to all the relations of this domain, 78.88% were obtained in the first three levels, and the percentage of correct relations for these levels exceeds 96%.

The percentage of correctly generated relation for the botany domain was 71.31%, where 64.97% of all the relations were found at the first level, and the success rate for this level exceeds 99%.

Despite a large number of investigations have been carried out for extracting lexical-semantic relations from dictionaries, most of them have been for English. Unfortunately, no published works were found that reported the automatic exploitation of lexical-semantic knowledge from these linguistic resources for Spanish, and to which we could compare our approach.

References

1. Potter, S.: A Survey of Knowledge Acquisition from Natural Language. AKT project report Task 1.1.2, Edinburgh, Scotland (2003)
2. Sanchez, S.-C., Lloréns, J., Morato, J., Hurtado, J.A.: Extracción Automática de Relaciones Semánticas. In: Proc. Conferencia Iberoamericana en Sistemas, Cibernética e Informática CИСCI 2003, July 2003, pp. 265–268 (2003)
3. Dolan, W., Vanderwende, L., Richardson, S.D.: Automatically Deriving Structured Knowledge Bases from On-Line Dictionaries. In: Proc. First Conference of the Pacific Association for Computational Linguistics, May 1993, pp. 5–14 (1993)
4. Fellbaum, C.: WordNet: An Electronic Lexical Database. Language, Speech and Communication Series. MIT Press - Bradford Books, Cambridge (1998), <http://mitpress.mit.edu>
5. Ortega, R.M.: Descubrimiento Automático de Hipónimos a Partir de Texto no Estructurado. M.S. Thesis, Instituto Nacional de Astrofísica, Óptica y Electrónica, Puebla, Mexico (December 2007)
6. Hearst, M.A.: Automatic Acquisition of Hyponyms from Large Text Corpora, Computer Science Division, University of California, Berkeley (1992)
7. Pantel, P., Pennacchiotti, M.: Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations. In: Proc. Conf. on Computational Linguistics / Association for Computational Linguistics (COLING/ACL), Sydney, Australia, pp. 113–120 (2006)
8. Cimiano, P.: Ontology Learning and Population from Text, Algorithms, Evaluation and Applications. Springer, New York (2006)
9. Alshawi, H.: Processing Dictionary Definitions with Phrasal Pattern Hierarchies. Computational Linguistics, 195–202 (1987)
10. Calzolari, N.: Detecting Patterns in a Lexical Data Base. In: Proc. 22nd Annual Meeting of the Association for Computational Linguistics (ACL), Stanford, California, pp. 170–173 (1984)
11. Hjelm, H., Buitelaar, P.: Multilingual Evidence Improves Clustering-based Taxonomy Extraction. In: Proc. 18th European Conference on Artificial Intelligence, ECAI 2008 (2008)
12. Diccionario enciclopédico LEXI-K, Mexico, Emblem Interactive (1997)
13. Mendoza, A.: Construcción de un Preprocesador de Consultas en Lenguaje Natural a una Base de Datos, M.S. Thesis, Instituto Tecnológico de Ciudad Madero (October 2004)

AVEDA: Statistical Tests for Finding Interesting Visualisations

Katharina Tschumitschew¹ and Frank Klawonn^{1,2}

¹ Department of Computer Science
University of Applied Sciences Braunschweig/Wolfenbuettel
Salzdahlumer Str. 46/48, D-38302 Wolfenbuettel, Germany

² Helmholtz Centre for Infection Research
Department for Cell Biology
Inhoffenstr. 7, D-38124 Braunschweig, Germany

Abstract. Visualisation is usually one of the first steps in handling any data analysis problem. Visualisations are an intuitive way to discover inconsistencies, outliers, dependencies, interesting patterns and peculiarities in the data. However, due to modern computer technology, a vast number of visualisation techniques is available nowadays. Even if only simple scatterplots, plotting pairs of variables against each other, are considered, the number of scatterplots is too large for high-dimensional data to visually inspect each scatterplot. In this paper, we propose a system architecture called AVEDA (Automatic Visual Exploratory Data Analysis) which computes a large number of visualisations, filters out those ones that might contain special patterns and shows only these interesting visualisations to the user. The filtering process for the visualisations is based on statistical tests and statistical measures.

1 Introduction

According to Tukey [1] “there is no excuse for failing to plot and look” when one wants to solve a data analysis problem. In the earlier days of data analysis, when computers were scarcely available, monitors were restricted to alpha-numeric displays, data visualisation was carried out manually, producing visualisations like bar charts, histograms, box plots, stem-and-leaves diagrams or simple scatterplots. This meant that only small data sets could be treated in this way and one could focus on one or at most two variables at the same time. Nowadays, computing and graphical displays allow fast computation of visualisations even for larger and high-dimensional data sets. This progress in computer technology enabled the application of more sophisticated visualisation techniques like multidimensional scaling (MDS) (see for instance [2]) or principal component analysis (PCA) (see for instance [3]), which need more computational power. But the progress in computer technology also led to the development of a vast number of new visualisation techniques for data analysis and data mining [4].

However, it is impossible for various reasons to check all possible visualisations individually for the following reasons.

- The number of different visualisations is too large, especially for high-dimensional data. Even if only scatterplots are considered, plotting pairs of variables against

each other, this would lead to $\binom{100}{2} = 4950$ plots when 100 attributes are available. The number of plots is already infinite when arbitrary projection planes are taken into account.

- Apart from the large number of visualisations, some techniques like MDS require high computational effort and are even unsuitable for larger data sets, although more efficient algorithms have been developed in the last years [5][6]. The same computational problems apply to nonlinear PCA techniques [7][8][9], where instead of a linear projection mapping as in standard PCA a nonlinear mapping is constructed.

Therefore, it can be very helpful to compute a larger number of possibly complex visualisations off-line and to present only those to the user which have interesting properties. Interesting properties are defined in terms of statistical tests and measures, for instance for correlation, independence, deviation from a certain distribution or multimodality. The visualisations and the tests to be applied can of course be preselected by the user. Only those visualisations will be presented to the user which render a significant value for the test or statistical measure.

Previous work on evaluating and selecting visualisations is briefly reviewed in section 2. In this paper we present a general system architecture for this approach which is described in section 3. We restrict our considerations to scatterplots which might be generated from simple projections, by MDS, PCA or other dimension reduction techniques. In section 4 we describe a selection of statistical tests that can be applied to to filter out interesting visualisations. We also discuss the often neglected problem of multiple testing. Illustrative examples are provided in section 5. In the final conclusions in section 6 we discuss perspectives for future work.

2 Projection Pursuit

The idea to select visualisations automatically based on suitable measures of interestingness was already proposed by Friedman and Tukey with their projection pursuit method [10]. Basically, after an affine transformation, the data set is projected onto (random) planes and the best projection w.r.t. a selected measure of interestingness is chosen. Very often, interestingness is defined as the deviation from a normal distribution according to the observation that most of the projections will resemble a normal distribution [11]. After estimating the probability density function f of the projection with a standard technique like kernel estimators, measures like the Friedman-Tukey index $E(f(x))$ [10] and the entropy index $E(\ln(f(x)))$ [12] or measures like the Legendre [13], the Hermite [14] or the natural Hermite index [15] indicating the relative, absolute and the expected squared error of f w.r.t. to the normal distribution. Also a χ^2 -test for normal distribution has been proposed [16].

However, the focus on deviations from a normal distribution is often too narrow. Other criteria might also be of interest as described in the following section.

3 Architecture

The architecture of our system AVEDA (Automatic Visual Exploratory Data Analysis) is illustrated in figure 1.

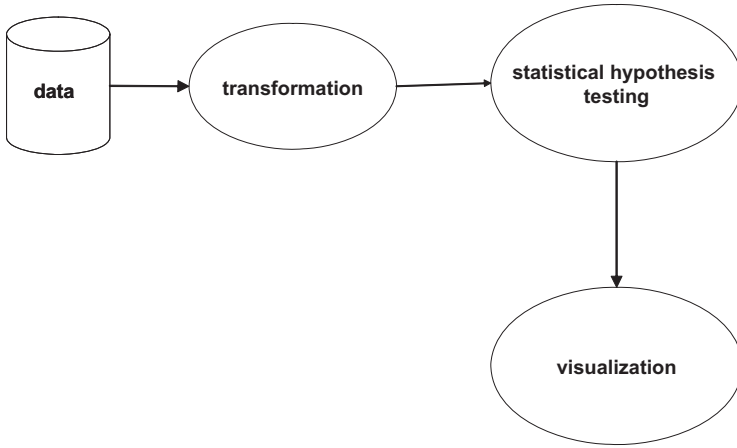


Fig. 1. AVEDA architecture

AVEDA provides an extensible set of visualisations and an extensible set of statistical tests. After the data have been provided, the types of visualisations to be applied are selected by the user as well as the statistical tests of interest. Then the data set X is transformed according to the visualisations, however, without generating the actual graphical displays immediately. This leads to a sequence $\{\tilde{X}_t\}$ of transformed data sets, in our case to two-dimensional representations suitable for scatter plots. The selected statistical tests $S = \{s_i\}$ are applied to these transformed data sets and only those scatter plots are generated for which the tests yield significant results.

Each statistical test s_i will select a subset of relevant visualisations \tilde{X}_v from all possible visualisations \tilde{X}_t . $s_i : \{\tilde{X}_t\} \rightarrow \{\tilde{X}_v\}$ where $\{\tilde{X}_v\} \subset \{\tilde{X}_t\}$. Only the visualisations $\{\tilde{X}_v\}$ will be presented to the user.

Here we consider only visualisations leading to two- or three-dimensional scatter-plot representations. In the simplest case, two or three attributes are plotted against each other. This corresponds to axes-parallel projections. Special linear projections, for instance onto the first two or three principal components can be considered or projections onto random planes or hyperplanes. But also nonlinear techniques like MDS or nonlinear PCA can be included. The statistical tests include for example

- tests for independence (χ^2 -test),
- tests for conditional independence,
- tests for structures in data (Hopkins index),
- tests for (multidimensional) outliers and
- tests for goodness of fit for given distributions, for instance uniform or normal distributions (two- or three-dimensional, depending on the chosen visualisation).

The set S of tests can be extended and the same applies to the transformation to generate the scatterplots.

4 Statistical Hypothesis Testing

There are two advantages of using statistical tests instead of arbitrary measures of interest. The p -value obtained from a test, i.e. the lowest error probability for rejecting the null hypothesis, has a clear interpretation and different tests, even if they are designed for different characteristics can be compared, which is not true for arbitrary indices.

One can simply rank the visualisations based on their assigned p -values, no matter from which test the p -value comes. The user can browse through the visualisation, starting with those with lowest p -values. It is, of course, also possible to choose a p -value or α -level in advance and to select only those visualisations that are considered to be significant w.r.t. this level. In this case, the problem of multiple testing should be taken into account, since not only one test for one visualisation is carried out, but many tests simultaneously. Therefore, in order to set the p -value or α -level correctly, methods like Bonferroni correction [17] or the further improved Bonferroni-Holm method [18] should be applied.

5 Examples

In this section, we demonstrate how AVEDA can be applied. First we consider an artificial data set $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^p$. The data originate from a uniform distribution over a chessboard-like structure within the m -dimensional unit hypercube, i.e. the unit hypercube is divided into 2^p subcubes and every other subcube does not contain any data. This data set is interesting in so far that all scatterplots plotting all pairs of attributes against each other will show no indication of any structure. Figure 2 shows the original data set for $p = 3$. The corresponding scatterplots are illustrated in figure 3.

The following transformations were applied to this data set: parallel projections as in figure 3, projection of the first two principal components and projections to random planes that are not necessarily parallel to the coordinate axes. As statistical tests or measures of interest, we applied the Hopkins index and the χ^2 -test for independency.

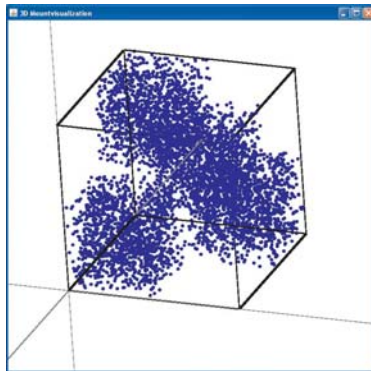


Fig. 2. Chessboard cube

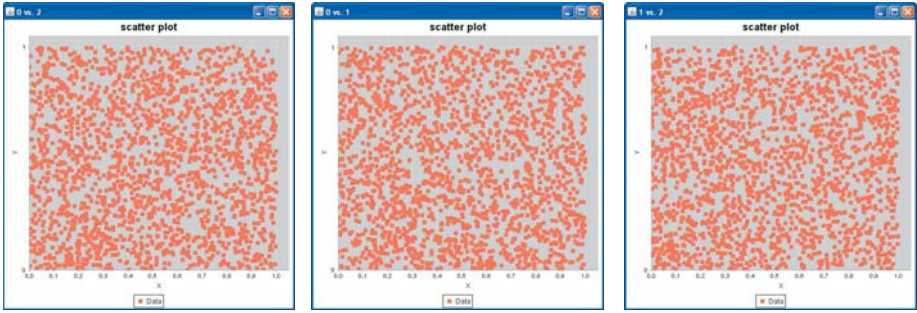


Fig. 3. Parallel projections for the chessboard cube

5.1 Test for Structures (Clusters) in Data

The Hopkins index [19] is a measure for the presence of structures in the form of clusters in a data set. For a given data set $X = \{x_1, \dots, x_n\} \subset \mathbb{R}^p$, the Hopkins index is computed in the following way. First a set $R = \{r_1, \dots, r_m\} \subset \mathbb{R}^p$ of m points from a uniform distribution over the convex hull of X and a set $S = \{s_1, \dots, s_m\} \subset \mathbb{R}^p$ of m randomly chosen points from the data set X where ($m \ll n$) are generated. Then the distances d_{r_1}, \dots, d_{r_m} and d_{s_1}, \dots, d_{s_m} defined by $d_{r_i} = \min_{x \in X} \{d(x, r_i)\}$ and $d_{s_i} = \min_{x \in X} \{d(x, s_i)\}$ are computed. The Hopkins index is defined as $h = \frac{\sum_{i=1}^m d_{r_i}^p}{\sum_{i=1}^m d_{r_i}^p + \sum_{i=1}^m d_{s_i}^p}$. It compares the distances d_{r_i} for the random points – considered as random cluster centres – with the distances d_{s_i} for the randomly chosen points from the data set. Under the null hypothesis H_0 that there is no structure in the data, the two sets of distances should have roughly the same distribution. In order to avoid random effects by an inappropriate choice of the sets R and S , the Hopkins index is usually computed for a number of random sets R and S and the average of these values is taken as the overall Hopkins index. The interpretation of the Hopkins index $0 \leq h \leq 1$ is as follows.

1. $h \approx 1$ means that the distances d_{r_i} are large in comparison to the distances d_{s_i} . This is caused by regions of higher data density (clusters) where data points are closer to each other than randomly chosen points in the convex hull of the data set.
2. $h \approx 0.5$ occurs when the distances d_{s_i} are similar to the distances d_{r_i} . This happens, when the random points in the convex hull of the data have roughly the same distribution as the data points, indicating that there is no structure in the data set.
3. $h \approx 0$ means that the distances d_{s_i} are large compared to the distances d_{r_i} , indicating a regular distribution of the data with approximately the same distances between them (for instance on a grid) and not a uniform distribution of the data.

Therefore, the Hopkins index can help to discover whether there is a tendency for clusters in the data ($h \gg 0.5$), the data follow more or less a random distribution ($h \approx 0.5$) or the data have a roughly regular underlying structure¹ ($h \ll 0.5$).

¹ This grid effect can be caused by limited precision measurements. For precision of 0.1, the data points lie automatically on the grid $(\frac{1}{10}\mathbb{Z}) \times (\frac{1}{10}\mathbb{Z})$.

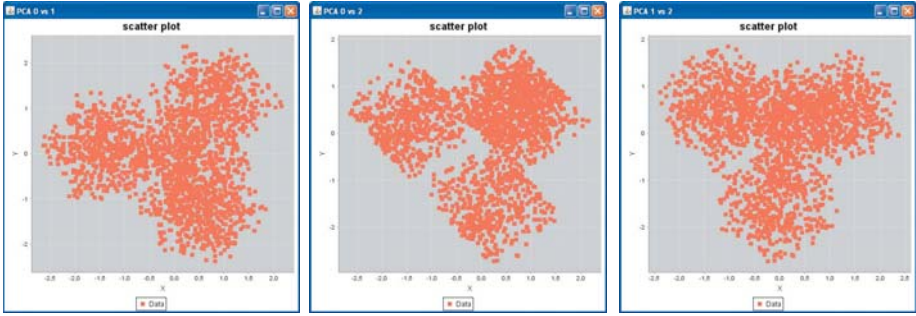


Fig. 4. Projections to pairs of principal components

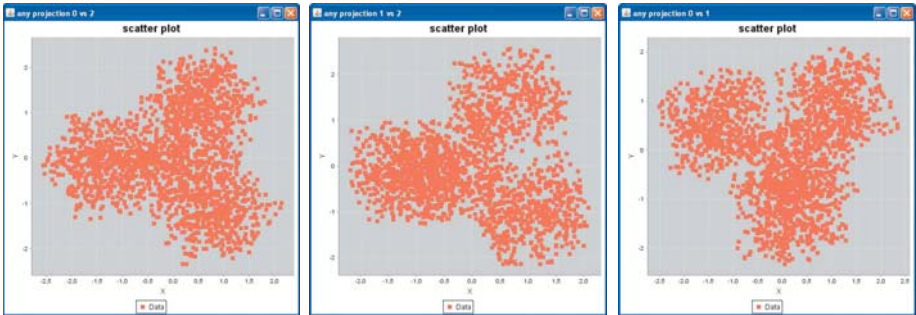


Fig. 5. Projections to random planes

For the artificial data set from figure 2, the Hopkins index has been computed for various projections. Only those visualisations are shown which have a Hopkins index larger than 0.8. For the parallel projections in figure 3, the Hopkins index was always smaller than 0.8 and none of these visualisations would be shown to the user. Projections to pairs of principal components were also computed as shown in figure 4. All these projections would be visualised, since their Hopkins index is larger than 0.8. Projections to random planes can also lead to a Hopkins index larger than 0.8, leading to visualisations as they are shown in figure 5.

5.2 χ^2 -Test

The χ^2 -test has various applications. The principal idea of the χ^2 -test is the comparison of two distributions. One can check whether two samples come from the same distribution, a single sample follows a given distribution or also whether two samples are independent. The null hypothesis H_0 for the χ^2 -test claims that the data follow a certain (cumulative) probability distribution $F(x)$. The distribution of the null hypothesis is then compared to the distribution of the data. The null hypothesis can for instance be a given distribution, for instance a uniform or a normal distribution, and the χ^2 -test can give an

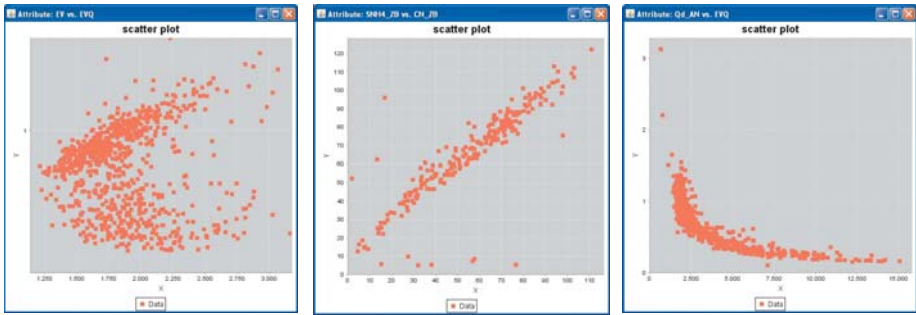


Fig. 6. Real data from a wastewater treatment plant

indication, whether the data strongly deviate from this expected distribution. For an independence test for two variables, the joint distribution of the sample is compared to the product of the marginal distributions. If these distributions differ significantly, this is an indication that the variables might not be independent.

We have applied the χ^2 -test for independence to real world data set from a wastewater treatment plant. Measurements for 59 attributes concerning chemical and physical properties of the wastewater were collected over a period of three years. Even simple projections to pairs of attributes would lead to $\binom{59}{2} = 1711$ scatterplots. The χ^2 -test for independence including correction for multiple testing reduced the visualisations to a still high number of 445 significant correlations. The visualisations with the most significant p -values are shown in figure 6. Of course, 445 visualisations would still be a too high number of visualisations to be checked. The reason here is that there strong interdependencies among the variables, many of them known, but not all of them. Such information can be used to check only those visualisations where an unsuspected dependence occurs.

6 Conclusions

In this paper, we have proposed a general architecture called AVEDA for selecting interesting visualisations from a larger number of visualisations based on statistical tests and measures. So far we have only considered diagrams in the form scatterplots induced by transformations of the data. Future work will include other visualisation methods like parallel coordinates as well. We will also introduce a framework that allows the user to specify known structural and dependence information, so that visualisations showing significant, but known patterns in the data are not selected as interesting. Another possible extension of AVEDA is to take labeled or classified data into account, preferring visualisations that separate classes as it is proposed in [20] based on a nearest neighbour classifier.

Acknowledgments. The authors would like to express their gratitude for the valuable comments of two anonymous reviewers that helped to improve the paper significantly.

References

1. Tukey, J.W.: Exploratory Data Analysis. Addison-Wesley, Reading (1977)
2. Borg, I., Groenen, P.: Modern Multidimensional Scaling: Theory and Applications. Springer, Berlin (1997)
3. Jolliffe, I.: Principal Component Analysis. Springer, New York (2002)
4. Soukup, T., Davidson, I.: Visual Data Mining: Techniques and Tools for Data Visualization and Mining. Wiley, New York (2002)
5. Morrison, A., Ross, G., Chalmers, M.: Fast multidimensional scaling through sampling, springs and interpolation. Information Visualization 2 (2003)
6. Rehm, F., Klawonn, F., Kruse, R.: MDS_{polar} – a new approach for dimension reduction to visualize high dimensional data. In: Famili, A.F., Kook, J.N., Peña, J.M., Siebes, A., Feelders, A. (eds.) IDA 2005. LNCS, vol. 3646, pp. 316–327. Springer, Heidelberg (2005)
7. Lowe, D., Tipping, M.: Feed-forward neural networks topographic mapping for exploratory data analysis. Neural Computing and Applications 4, 83–95 (1996)
8. Scholz, M., Kaplan, F., Guy, C., Kopka, J., Selbig, J.: Non-linear pca: A missing data approach. Bioinformatics 21, 3887–3895 (2005)
9. Kolodyazhniy, V., Klawonn, F., Tschumitschew, K.: Neuro-fuzzy model for dimensionality reduction and its application. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 15, 571–593 (2007)
10. Friedman, J., Tukey, J.: A projection pursuit algorithm for exploratory data analysis. IEEE Transactions on Computers C-23, 881–890 (1974)
11. Diaconis, P., Freedman, D.: Asymptotics of graphical projection pursuit. The Annals of Statistics 17, 793–815 (1989)
12. Huber, P.: Projection pursuit. The Annals of Statistics 13, 435–475 (1985)
13. Friedman, J.: Exploratory projection pursuit. Journal of the American Statistical Assoc. 82, 249–266 (1987)
14. Hall, P.: On polynomial-based projection indices for exploratory projection pursuit. The Annals of Statistics 17, 589–605 (1989)
15. Cook, D., Buja, A., Cabrera, J.: Projection pursuit indices based on orthonormal function expansion. Journal of Computational and Graphical Statistics 2, 225–250 (1993)
16. Posse, C.: Projection pursuit exploratory data analysis. Computational Statistics and Data Analysis 20, 669–687 (1995)
17. Shaffer, J.P.: Multiple hypothesis testing. Ann. Rev. Psych 46, 561–584 (1995)
18. Holm, S.: A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics 6, 65–70 (1979)
19. Hopkins, B.: A new method of determining the type of distribution of plant individuals. Annals of Botany 18, 213–226 (1954)
20. Leban, G., Bratko, I., Petrovic, U., Curk, T., Zupan, B.: VizRank: Finding informative data projections in functional genomics by machine learning. Bioinformatics 21, 413–414 (2005)

Degree of Association between Documents Using Association Mechanism

Hirokazu Watabe, Eriko Yoshimura, and Seiji Tsuchiya

Dept. of Intelligent Information Engineering and Science, Doshisha University,
Kyo-Tanabe, Kyoto, 610-0394, Japan
hwatabe@mail.doshisha.ac.jp, eyoshimura@indy.doshisha.ac.jp,
stsuchiy@mail.doshisha.ac.jp

Abstract. This paper proposes a method that quantifies the similarity between documents based on the level of relevance among terms in order to deliver a search that captures the meaning of documents. More specifically, this paper proposes a method that uses a concept-base to look for relevance among different terms and calculates the degree of association between documents using the Earth Mover's Distance. When the proposed methods were subjected to comparison tests with other methods using the NTCIR3-WEB, they achieved good results.

Keywords: Document search, Concept-Base, Earth Mover's Distance, Degree of association.

1 Introduction

With the advance of computers and networks, the quantity of electronic documents and other forms of information is now becoming enormous. Under these circumstances, a method for strictly selecting and providing only critical portions of information is essential.

This study proposes employing a search method that uses a concept base that defines the semantic features of the search terms [1] in order to create a search that captures the meaning within a document. Using a concept base allows a search that captures a term's meaning, which is unlike a search method that only uses a term's notation. In other words, it is a method that is able to quantify the semantic closeness of terms without being influenced by the notational variations of the terms the user enters to perform the search. Specifically, definitional relevance between terms is calculated as values from 0 to 1 by the concept base. What we are proposing is a method that identifies the degree of similarity between the search request and the search target based on that value by using the Earth Mover's Distance (EMD) [2], which is a distance scale that is drawing attention in fields such as image searching.

As a related work, Vector Space Model [4] and Okapi BM25 [5], which employ notation, have been proposed. And a method that uses WordNet [6], a systematically organized dictionary, to define the distance between terms and EMD to define the degree of similarity between documents has been proposed [7]. This allows information searches that focus on the semantic relevance of terms.

2 Association Mechanisms

To understand the contents of a document, we use Concept Base [1], which expresses the semantic characteristics of a word with a word and a weight, and also the method of calculating the degree of association, which numerically calculates the semantic relationship between words.

2.1 Concept-Base

A Concept is defined as the following equation.

$$A = \{(a_1, w_1), (a_2, w_2), \dots, (a_m, w_m)\}$$

Where, A is named the concept label (a word), a_i is called the attribute, and w_i is called the weight. The Concept-Base is a set of concepts and is a knowledge base of approximately 120,000 concepts, and it is constructed from Japanese dictionaries and other information. Fig. 1 shows an example of the Concept-base.

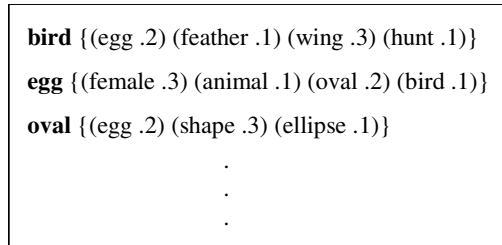


Fig. 1. Example of the concept-base

2.2 Degree of Association

Each concept is defined as a set of attributes and each attribute is also a concept as described above. In this paper, a method to derive the degree of association between concepts using up to second order attributes is used. The value of the degree of association ranges from 0.0 to 1.0.

The quantifying of the degree of association between terms using the concept base basically uses the deployment results of the term's meaning and expresses that as a number. Since that value will vary depending on up to what order of the attribute is to deploy to and what attributes are used, the problem becomes how to perform the calculation in response to the situation. For this reason, this study makes use of two different methods. The degree of match, which is employed up to the primary attribute of the concept base, is used to quantify the relevance between terms so that the similarity between documents can be found. The degree of association calculation, which is employed up to the secondary attribute of the concept base, is used to quantify the relevance of the terms in the automatic expansion method of the concept base. It has been reported that the method of using up to the secondary attribute is the most effective for quantifying the relevance between terms that use the concept base. If deployment takes place only up to the primary attribute, the quantification of relevance between weakly related concepts

cannot take place. Additionally, if deployment takes place up to the tertiary attribute, terms that are far removed from the concept can become attributes, causing a decrease in accuracy since they function as noise. The level of agreement that deploys only up to the primary attribute is used in this study to quantify the relevance between terms in order to determine the degree of similarity between the document and the concept. If we liken a document to a concept, then an index term becomes the primary attribute and the attribute of the index term becomes the secondary attribute. In other words, if attributes up to the secondary attribute of the index term are deployed, and if the document is considered a concept, that means attributes up to the tertiary attribute are deployed. This increases noise as terms that are far removed from the concept (document) end up being used for the calculation.

Calculating the Degree of Match

Let us assume that the primary attributes for any concepts A and B are a_i and b_j , respectively, and that their corresponding weightings are u_i and v_j . The number of attributes for concepts A and B shall be L attributes and M attributes ($L \leq M$).

$$A = \{ (a_i, u_i) \mid i = 1 \sim L \}$$

$$B = \{ (b_j, v_j) \mid j = 1 \sim M \}$$

In this situation, the degree of match for concepts A and B , $Match(A,B)$, can be defined in the following equation.

$$Match(A,B) = \sum_{a_i=b_j} \min(u_i, v_j)$$

Where, the total sum of the weightings of each concept must be normalized to 1. If there is an attribute that becomes $a_i = b_j$ (concepts A and B have a common attribute), relative to attributes a_i and b_j of concepts A and B , the common portions of the weightings of the common attributes. This means that only the smallest portions of the weightings will effectively be considered to be in match, and their total shall be the degree of match.

Calculating the Degree of Association

To find the degree of association, the relevance between two concepts is calculated as a numeric value based on the value found by calculating the degree of match of the secondary attributes of the concepts. More specifically, of the two concepts to be calculated, let the one with the smallest number of primary attributes be A ($L \leq M$) and the primary attributes of concept A be the criteria.

$$A = \{ (a_1, u_1), \dots, (a_i, u_i), \dots, (a_L, u_L) \}$$

After that, rearrange the primary attributes of concept B so that the product of the degree of match, $Match(a_i, b_{xi})$, with each primary attribute of concept A is at a maximum.

$$B_x = \{ (b_{x1}, v_{x1}), \dots, (b_{xi}, v_{xi}), \dots, (b_{xL}, v_{xL}) \}$$

This will determine a correlating set for the primary attributes of concept A and the primary attributes of concept B . The primary attributes of concept B that exceed the correlation will be ignored. (There will be L attribute combinations at this time.)

However, if some of the primary attributes are in agreement, meaning that their concept notations are the same ($a_i=b_j$), they will be handled separately. This is because there are about 120,000 concept notations in the concept base and the matching of attributes is considered to be rare. As a result, by handling the matching attributes separately, they are valued more highly when they match. More specifically, the size of the corresponding attribute weightings u_i and v_j will be aligned in the direction of the smallest weighting. When this takes place, the value of the attribute with the smallest weighting will be subtracted from the attribute with the largest weighting and once again be correlated to the other attributes. For example, if $a_i=b_j$ and $u_i=v_j+\alpha$, then the correlation would be between (a_i, v_j) and (b_j, v_j) , and (a_i, α) would once again be correlated to the other attributes. Let us assume the number of attribute combinations determined and correlated in this manner is T . The degree of association between concepts A and B in this case, $DoA(A,B)$, is defined in equation below.

$$DoA(A,B) = \sum_{i=1}^T \{Match(a_i, b_{xi}) \times (u_i + v_{xi}) \times (\min(u_i, v_{xi}) / \max(u_i, v_{xi})) / 2\}$$

The value of the degree of association expresses the strength of the relevance between concepts as a continuous value between 0 and 1. The closer the value is to 1, the stronger the relevance.

3 Degree of Association between Documents Using EMD

When seeking the degree of similarity between a search request document and a search target document, no matter how accurately the relevance between terms can be defined, if the calculation cannot take place based on the values, it will be impossible to find the precise degree of similarity between the documents. A variety of methods can be used for the calculation. For instance, one method would be to perform the calculation by correlating the terms in order beginning from the highest degree of relevance between the terms. A method that involves a one-to-one correlation can only correlate to the smallest number of terms between the search request and the search target. For example, if the search request has three terms and the search target has 100 terms, 97 of the search target terms will not be subjected to calculation. Furthermore, it is believed that when performing the actual search, users will not enter many terms in the search request, so the assumption is that there will be a large difference in the number of terms in the search request and search target. Therefore, it is necessary to consider the importance of terms in the text and the relevance between them and to be flexible in handling M relative to N .

For this reason, the EMD [2], which has been drawing attention in the field of similar imagery searching, has been employed in this study as a method that calculates the degree of similarity between documents. The EMD is an algorithm that seeks the optimal solution for transportation costs in a transportation problem. As a result, if the weighting between the demand point and the supply point and the distance between these points are defined, it can be used to solve any type of problem. By employing the EMD and taking the weighting of terms and the relevance between terms into consideration, correlation can be flexible and the degree of similarity between sentences can be found.

3.1 What Is the EMD?

The EMD is a distance scale that calculates by means of the Hitchcock transportation problem, which is one type of linear programming problem. Given two discrete distributions, it is defined as the minimum cost of converting one distribution to the other distribution. The transportation problem is the problem of solving transportation from the supply point to the demand point in order to satisfy the demand at the demand point at minimum cost.

When seeking the EMD, the two distributions are expressed as sets that have been assigned element weightings. If one of the distributions P is expressed as a set, the expression becomes, $P = \{(p_1, w_{p1}), \dots, (p_m, w_{pm})\}$. Distribution P is currently expressed as having m number of characteristics. p_i represents the characteristics, while w_{p_i} represents the weighting of the characteristics. In like manner, if the other distribution Q is expressed as a set, the expression becomes, $Q = \{(q_1, w_{q1}), \dots, (q_n, w_{qn})\}$. As for the EMD calculation, even if the number of characteristics for both distributions differs, it has a characteristic that allows the calculation to take place. Let us assume that the distance between p_i and q_j is d_{ij} and the distance between all features is $D = [d_{ij}]$. If we assume the amount of transportation from p_i to q_j to be f_{ij} , the total amount of transportation becomes $F = [f_{ij}]$. Here, we will find the amount of transportation F , which creates the minimum cost function shown in the following equation, and calculate the EMD.

$$WORK(P, Q, F) = \sum_{i=1}^n \sum_{j=1}^m d_{ij} f_{ij}$$

However, when minimizing the above cost function, the following restrictions must be satisfied.

$$f_{ij} \geq 0, 1 \leq i \leq m, 1 \leq j \leq n \quad (1)$$

$$\sum_{j=1}^n f_{ij} \leq w_{p_i}, 1 \leq i \leq m \quad (2)$$

$$\sum_{i=1}^m f_{ij} \leq w_{q_j}, 1 \leq j \leq n \quad (3)$$

$$\sum_{i=1}^m \sum_{j=1}^n f_{ij} = \min \left(\sum_{i=1}^m w_{p_i}, \sum_{j=1}^n w_{q_j} \right) \quad (4)$$

In this case, we know that the amount of transportation in equation 1 is positive and we also know that transportation goes one way, from p_i to q_j . Equation 2 indicates that transportation cannot take place above the weighting of the transportation source p_i . Equation 3 indicates that acceptance cannot take place above the weighting of the transportation destination q_j . Finally, equation 4 indicates the upper limit of the total amount of transportation and is limited by the smaller of the sum total of either the transportation destination or transportation source. The EMD between distributions P and Q can be found as indicated below by using the optimal total amount of transportation F found under the limitations indicated above.

$$EMD(P, Q) = \frac{\sum_{i=1}^m \sum_{j=1}^n d_{ij} f_{ij}}{\sum_{i=1}^m \sum_{j=1}^n f_{ij}} \tag{5}$$

The reason the optimal cost function $WORK(P, Q, F)$ is used as is here as the EMD is that the cost function depends on the sum total of the weighting of either the transportation source or the transportation destination. So, that influence will be eliminated by normalization.

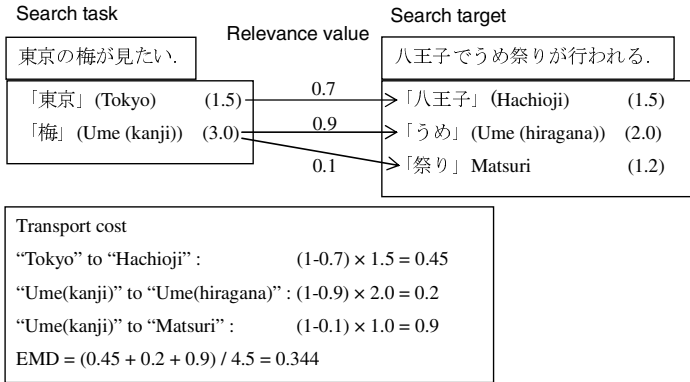


Fig. 2. Examples of Applying EMD to Document Search

3.2 Applying EMD to Document Search

Figure 2 shows examples of applying EMD to document search. To apply EMD to document search, the demand point and supply point, demand volume and supply volume, and the distance between each demand point and supply point must be defined. For the demand point, the index term for the search task is assigned, while for the supply point, the index term for the search result is assigned. The demand volume and supply volume each use the tf-idf weighting, which concerns index terms. The distance between the demand point and supply point can be considered to be the relevance between index terms and, thus, can be found in the proposed methodology by a degree of association that uses the concept base. Since the value of the degree of association will be larger as the relevance increases, it will be converted into a value in which the degree of association value will be subtracted from 1. The calculation of EMD is located at the bottom of Figure 2. The reason that the amount of transportation between "ume" and "matsuri" is 1 is because a weighting of 2 was transported from "ume (kanji)" to "ume (hiragana)" and the excess weighting of "ume (kanji)," 1, was transported to "matsuri." The weighting is transported in this manner to terms with a high degree of relevance and the transportation will take place until the supply volume disappears or the demand volume is satisfied. In this way, a flexible M versus N that considers relevance and weighting between index terms is possible. As a characteristic of the EMD, if the value of the distance between index terms is from 0 to 1, then EMD

also becomes a value from 0 to 1. Additionally, if there is similarity between documents, the value falls, and if there is a lack of similarity, the value rises. Thus, document retrieval is realized by presenting documents to the user in sequence beginning with documents with low values.

4 Experiment and Evaluation

In order to verify the validity of the proposed method, which focuses on the relevance between terms, we used the information search test collection NTCIR3-WEB [3] to make a comparison with another method that employs notation. As a means of comparison, we employed vector space model [4] and Okapi BM25 [5].

4.1 Evaluation Method

At this stage, we conducted the evaluation test by using 41 search tasks and 10,000 documents, including matching documents and randomly selected documents. In addition, a matching document list was used and, for each search result of each document, there were four levels of matching: H (high level match), A (match), B (partial match), and C (no match). The evaluation took place as described below.

Level 1: H and A match.

Level 2: H and A and B match.

For each search task, we sought a score for all 10,000 search targets and rearranged them in the order of the score. We then referenced the matching document list, checked the sequence of the matching documents and made an evaluation.

4.2 Evaluation Indicator

As an evaluation indicator, for each search result, we used the average precision (AP), mean average precision (MAP), and a recall rate-accuracy graph. The AP for the search task was defined as shown in equation 6 below. Initially, we assume the variable z_i to indicate that a document in sequence position i is a match and give it a 1, and if not, give it a 0. S is the sum total of the matching documents, while n is the number of documents output.

$$AP = \frac{1}{S} \sum_{i=1}^n \frac{z_i}{i} \left(1 + \sum_{k=1}^{i-1} z_k \right) \quad (6)$$

The mean of the average precision (MAP) is an average of the average accuracy for all of the search results and is found as indicated in equation 7. Specifically, if the search result is K cases and we notate the average precision of the system as AP_h , we get ($h = 1, \dots, K$), and the average of this is equivalent to the MAP, as shown in the following equation.

$$MAP = \frac{1}{K} \sum_{h=1}^K AP_h \quad (7)$$

4.3 Evaluation Results

Table 1 shows the mean of the average precision (MAP). It shows that at Level 1 the accuracy of the proposed method is 20.2% better than the vector space model, 20.0% better than Okapi BM25. At Level 2, the accuracy of the proposed method is 8.0% better than the vector space model, 12.4% better than Okapi BM25.

Table 1. MAP (Mean Average Precision)

Method	MAP(Level 1)	MAP(Level 2)
VSM	0.4305	0.5793
Okapi BM25	0.4311	0.5569
Proposed method	0.5173	0.6259

5 Conclusions

This paper proposes a method of calculating the degree of association between documents by defining the relevance between index terms through a concept base and finding the similarity between documents using the EMD. The effectiveness of this method has been verified through the use of NTCIR3-WEB, the Web search evaluation test collection. Compared to other methods that rely on notation, the results have been good and the effectiveness of this method, which focuses on the relevance between terms, has been confirmed.

Acknowledgements

This research has been partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (Young Scientists (B), 21700241).

References

- [1] Okumura, N., Yoshimura, E., Watabe, H., Kawaoka, T.: An Association Method Using Concept-Base. In: Apolloni, B., Howlett, R.J., Jain, L. (eds.) KES 2007, Part I. LNCS (LNAD), vol. 4692, pp. 604–611. Springer, Heidelberg (2007)
- [2] Rubner, Y., Tomasi, C., Guibas, L.: The earth mover's distance as a metric for image retrieval. *Int. J. Comput. Vision* 40, 99–121 (2000)
- [3] <http://research.nii.ac.jp/ntcir/ntcir-ws3/ws-ja.html>
- [4] Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. *Communications of the ACM* 18(3), 613–620 (1975)
- [5] Robertson, S.E., Walker, S., Jones, S., Beaulieu, M., Gatford, M.: Okapi at TREC-3. In: *Proceeding of the 3rd Text Retrieval Conference*, pp. 109–126 (1995)
- [6] Miller, G.A.: "WordNet": A lexical database for English". *Commun. ACM* 38(11), 39–41 (1995)
- [7] Wan, X., Peng, Y.: The Earth Mover's Distance as a Semantic Measure for Document Similarity. In: *Proc. of 14th ACM international conference on Information and knowledge management*, pp. 301–302 (2006)

Parallel Method for Mining High Utility Itemsets from Vertically Partitioned Distributed Databases

Bay Vo¹, Huy Nguyen², Tu Bao Ho³, and Bac Le⁴

¹ Faculty of Information Technology, Ho Chi Minh City University of Technology, Vietnam

² Faculty of Information Technology, Saigon University, Ho Chi Minh, Vietnam

³ School of Knowledge Science, Japan Advanced Institute of Science and Technology, Japan

⁴ Faculty of Information Technology, University of Science, Ho Chi Minh, Vietnam

vdbay@hcmhutech.edu.vn, nghuy@sgu.edu.vn, bao@jast.ac.jp,
lhbac@fit.hcmuns.edu.vn

Abstract. Mining high utility itemsets (HUIs) has been developing in recent years. However, the methods of mining from distributed databases have not mentioned yet. In this paper, we present a parallel method for mining HUIs in vertically partitioned distributed databases. We use WIT-tree structure to store local database on each site for parallel mining HUIs. The item i^{th} in each SlaverSite is only sent to MasterSite if its Transaction-Weighted Utilization (TWU) satisfies minutility (minutil), and MasterSite only mines HUIs which exist at least on 2 sites. Besides, the parallel performance is also interesting because it reduces the waiting time of attended sites. Thus, the mining time is reduced more significant than that in mining from centralized database.

Keywords: Centralized database, high utility itemset, parallel, vertically partitioned distributed databases, WIT-tree.

1 Introduction

Mining high utility itemsets is the general form of frequent itemsets (FIs) mining [1]. It finds the itemsets which has high utility value from database. However, it is not like FIs mining, HUI does not satisfy the Apriori property. That is, the subset of a HUI is not likely a HUI. Therefore, we can not use fully algorithms of FIs for mining HUIs.

In 2004, H. Yao, H. J. Hamilton [14] proposed the model of mining HUIs. They proposed the UMining and the UMining_H (UMining with heuristic) algorithms to find HUIs [15]. Recently, some algorithms which based on TWU have been developed [3, 4, 6, 7]. The Two-Phase algorithm was firstly proposed by Y. Liu et al [7]. After that, some efficient algorithms were proposed [3, 4], they based on the methods which do not generate candidates. In [6], authors proposed WIT-tree, a new data structure, and an efficient algorithm for mining HUIs.

Although there are many algorithms for mining HUIs, but there is not any parallel method in distributed databases.

In this paper, we propose the parallel method for mining HUIs from vertically partitioned distributed databases. The main contributions of this paper are as follows:

- We define problem for mining HUIs from vertically partitioned distributed databases.
- We propose the parallel method for mining HUIs that scans local databases one time and need not integrate the databases in all sites together. It mines HUIs in local databases at SlaverSites, and MasterSite mines HUIs which are contained in at least two sites only. Each item in SlaverSite is only sent to MasterSite if its *twu* satisfies *minutil*. Therefore, we save a lot of time in communication between MasterSite and SlaverSites.

2 Related Works

In recent years, many HUIs algorithms have been proposed [3, 4, 6, 7, 14, 15, 16, 17]. The usefulness of an itemset is characterized as a utility constraint. That is, an itemset is interesting to the user if its utility only satisfies *minutil*. The usefulness of an itemset is computed by objective and subjective values of items. The objective value of an item, denoted x_{pq} , is the value of an attribute associated with an item i_p in a transaction t_q . The subjective value of an item, denoted y_p , is a real number assigned by the user such that for any two items i_p and i_q , y_p is greater than y_q if the user prefers item i_p to item i_q . The utility based itemset mining problem discovers the set of all high utility itemsets, i.e., HUIs = $\{S \mid S \subseteq I, u(S) \geq \text{minutil}\}$.

$$u(S) = \sum_{i_p \in S} \sum_{t_q \in T_S} f(x_{pq}, y_p) \tag{1}$$

where $f(x_{pq}, y_p) = x_{pq} \cdot y_p$, and T_S is the set of transactions that contains itemset S .

2.1 Estimated Utility Value Method

H. Yao, H. J. Hamilton [14] reduced search space by pruning candidates based on estimated utility value. Based on the utility upper bound $b(S^k)$, H. Yao, H. J. Hamilton proposed UMining [15] algorithm for mining all high utility itemsets.

2.2 Transaction-Weighted Utilization Value Method

Y. Liu et al [7] reduced search space by pruning candidates based on *twu* values. The utility of an itemsets S is always less than or equals the *twu* value of S ,

$$twu(S) = tu(T_S) = \sum_{t_q \in T_S} tu(t_q) = \sum_{t_q \in T_S} \sum_{i_p \in t_q} f(x_{pq}, y_p) \tag{2}$$

$$u(S) = \sum_{t_q \in T_S} \sum_{i_p \in S} f(x_{pq}, y_p) \leq \sum_{t_q \in T_S} \sum_{i_p \in t_q} f(x_{pq}, y_p) = twu(S) \tag{3}$$

the *twu* values satisfy the Apriori property [1].

A. Erwin et al [3, 4] proposed the efficient algorithms using the pattern growth approach. They have developed a new compact data representation named *Compressed Utility Pattern tree (CUP-tree)* which extends the *CFP-tree* [12] for mining HUIs, and a new algorithm named *CTU-PRO* for mining HUIs. The concept of TWU is used for pruning the search space in *CTU-PRO*, but it must re-scan the database to determine the actual utility of high *twu* itemsets. The algorithm creates a *CUP-Tree* named *GlobalCUP-Tree* from the transactions database after the first time of identifying the individual high *twu* items. For each high *twu* item, a smaller projection-tree called *LocalCUP-Tree* is extracted from the *GlobalCUP-tree* for mining all HUIs beginning with that item as prefix.

B. Le et al [6] proposed WIT-tree data structure and the algorithm for mining HUIs (TWU-Mining algorithm). We recognize that it is suitable for parallel mining HUIs in vertically partitioned distributed databases because TWU-Mining is vertical-based method (database is transformed into vertical format) and scans database one time.

2.2.1 WIT-Tree Data Structure

a) **Vertex:** Includes 3 fields

{X: an itemset; Tidset(X): the set of transaction IDs contains X; And *twu*: The sum of transaction-weighted utility of X.}. A vertex is denoted $X \times_{twu(X)} Tidset(X)$.

The *twu* value of X is computed by summing all *twu* values of transactions which their *tids* are contained in Tidset. Thus, the computing of *twu*(X) and of *u*(X) will be done quickly by using Tidset.

b) **Arc:** Connecting the vertex at k^{th} level (called X) with the vertex at $(k+1)^{th}$ (called Y) in which $X \equiv_{\theta_k} Y$ (X and Y have the same k-prefix) [18].

Example: Consider the following database

Table 1. Objective values table

Item \ TID	A	B	C	D	E
T ₁	0	0	16	0	1
T ₂	0	12	0	2	1
T ₃	2	0	1	0	1
T ₄	1	0	0	2	1
T ₅	0	0	4	0	2
T ₆	1	2	0	0	0
T ₇	0	20	0	2	1
T ₈	3	0	25	6	1
T ₉	1	2	0	0	0
T ₁₀	0	12	2	0	2

Table 2. Subjective values table

Item	Benefit
A	3
B	5
C	1
D	3
E	5

We have WIT-tree as in Figure 1:

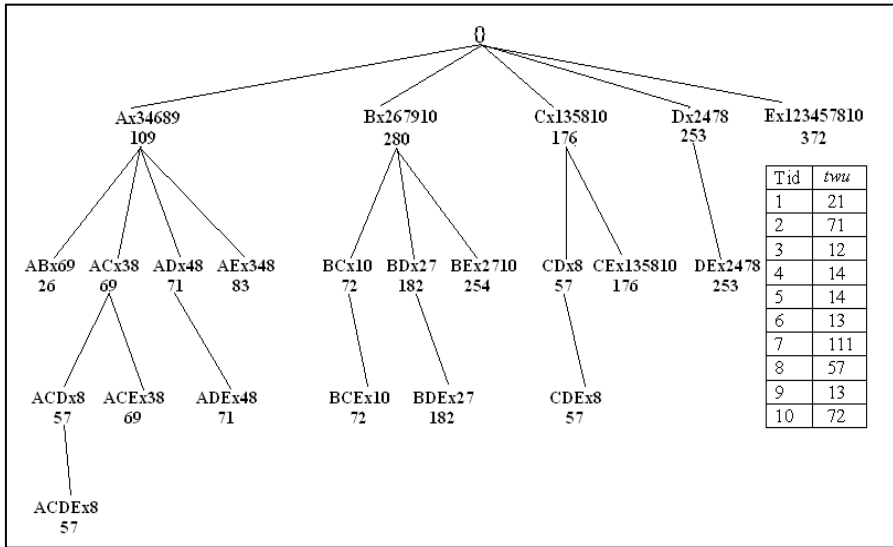


Fig. 1. Search tree using WIT-tree

2.2.2 TWU-Mining Algorithm

The TWU-Mining algorithm based on WIT-tree to mine HUIs. More details, we can see in [6].

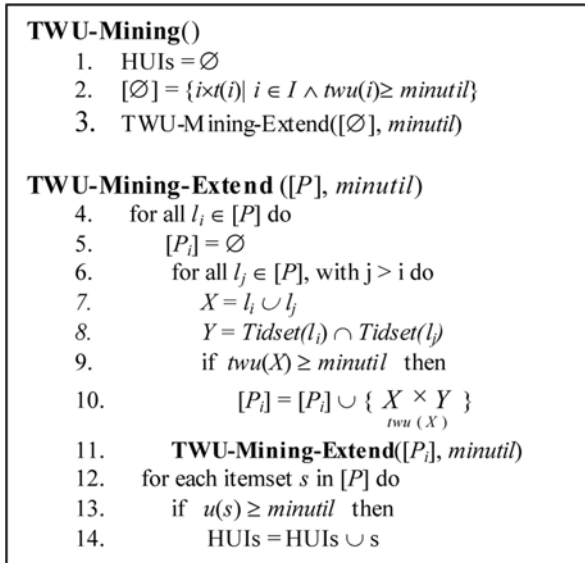


Fig. 2. TWU-Mining algorithm

2.3 Distributed Data Mining

There are many methods for mining in distributed databases such as: association rules [2, 10, 13, 19], classification [8, 9]. In [11], M. Serazi et al proposed an API that is transparent distributed vertical database. However, it can use for distributed databases, but can not be suitable for mining HUIs.

According to our knowledge, the parallel method for mining HUIs from distributed databases has not been proposed.

3 Parallel Method for Mining HUIs

3.1 Problem

A supermarket sold n items $I = \{i_1, i_2, \dots, i_n\}$, because of the specialization necessary, the supermarket needs to store information of customers in k computers (k sites), i.e., each site stores information of items (set of products). We can formularize as follows:

Database D is divided into k sites $\{D_1, D_2, \dots, D_k\}$, where D_j contains the set of items $I_j = \{i_{j_1}, i_{j_2}, \dots, i_{j_v}\}$ (v is number of items in sites D_j), the transactions in D_j only

contain the item that is in I_j . Assume that $I_i \cap I_j = \emptyset, \forall i \neq j$ and $\bigcup_{j=1}^k I_j = I$. When each

transaction is created, the new transaction ID, the items which are bought and the number of items are updated in the corresponding sites. Therefore, it is not being the centralized database, and makes the supermarket be easy to manage and to be not overloaded in case of huge amount of data.

The problem is how to mine HUIs from databases of many sites which do not integrate them together (database is very huge in centralized)?

Example: Consider the database given in Table 1, assume that it is distributed to 2 sites as follows:

Site 1:

Table 3. Objective values table

	A	B	C
T ₁	0	0	16
T ₂	0	12	0
T ₃	2	0	1
T ₄	1	0	0
T ₅	0	0	4
T ₆	1	2	0
T ₇	0	20	0
T ₈	3	0	25
T ₉	1	2	0
T ₁₀	0	12	2

Table 4. Subjective values table

Item	Benefit
A	3
B	5
C	1

Site 2: **Table 5.** Objective values table

	D	E
T ₁	0	1
T ₂	2	1
T ₃	0	1
T ₄	2	1
T ₅	0	2
T ₇	2	1
T ₈	6	1
T ₁₀	0	2

Table 6. Subjective values table

Item	Benefit
D	3
E	5

Table 1 is distributed into Table 3 and Table 5. Table 4 and Table 6 are subjective values that are corresponding to Table 3 and Table 5.

3.2 DTWU-Mining

Because the local HUIs are mined and sent to MasterSite from all SlaverSites, MasterSite only mines HUIs that its itemset appears at least in two SlaverSites. Therefore, we need to expand TWU-Mining for this purpose. When SlaverSites send information to MasterSite, we add the 4th field which is the group that indicates what SlaverSite contains that item. While we join 2 vertexes at level 1, we only check whether they are the same group or not. From the level 2, DTWU-Mining is the same as TWU-Mining (see fig. 3).

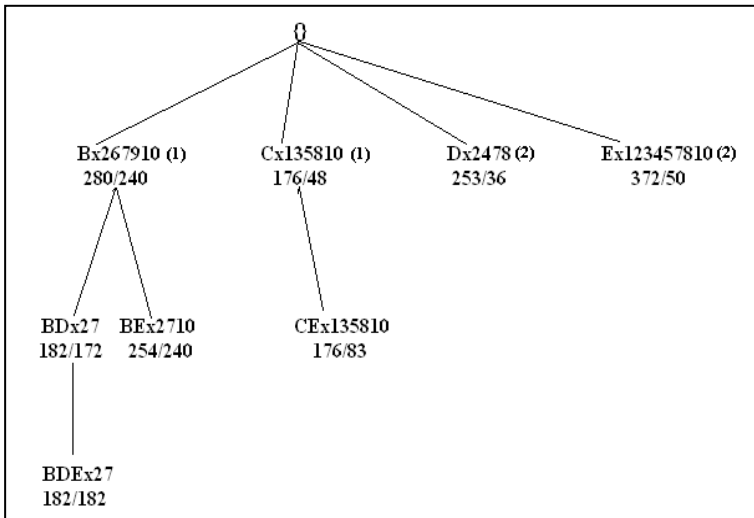


Fig. 3. WIT-tree for DTWU-Mining with *minutil* = 130

Because $\{B, C\}$, $\{D, E\}$ are the same group so they can not join together.

The pairs $\{B, D\}$, $\{B, E\}$, $\{C, D\}$, $\{C, E\}$ belong to 2 different groups so they can join together and become BD , BE , CD , CE . However, because of $twu(CD) < minutil$, CD is not generated in WIT-tree.

3.3 Parallel Algorithm

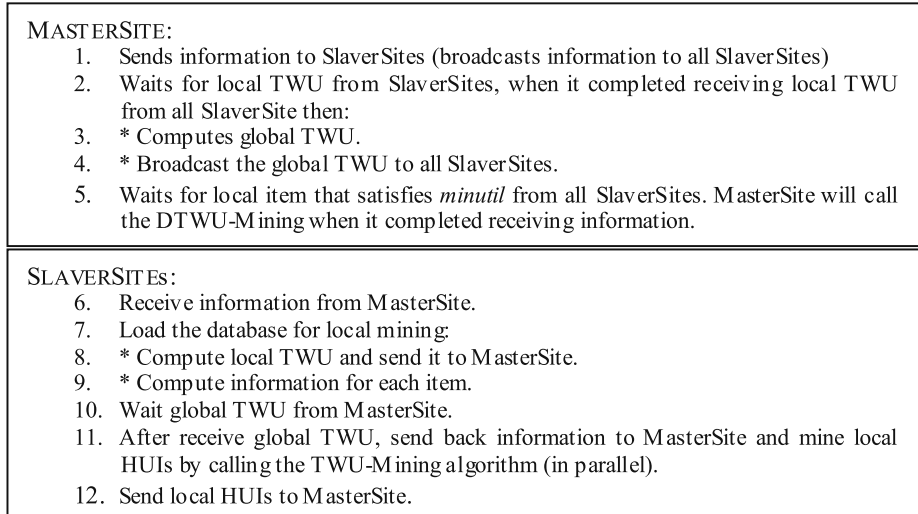


Fig. 4. General model for parallel mining HUIs from vertically partitioned distributed databases

MasterSite: First of all, MasterSite broadcast the request to all SlaverSites (name of databases and $minutil$) and waits local TWU from all SlaverSites (line 2). When MasterSite receives enough TWU from all SlaverSites, it computes and sends global TWU to SlaverSites (line 3-4). After that, it will call DTWU-Mining when it receives item which its twu satisfies $minutil$ (line 5).

SlaverSite: When the SlaverSite connects to the MasterSite, it will receive the necessary information from the MasterSite. After loading local database (line 7), the SlaverSite computes the local TWU and sends to MasterSite (line 8). After that, it computes the necessary information to send to MaterSite (line 9) and waits global TWU from MasterSite (line 10). When it receives TWU from MasterSite, it will send information of the item which its twu satisfies $minutil$ to MasterSite and mine local HUIs (line 11). Finally, it sends local HUIs to MasterSite (line 12).

Table 7, 8, 9, 10 illustrate the computing results in SlaverSites and information that collect from MasterSite for the mining process (with $minutil = 130$).

SlaverSite1: **Table 7.** Tidset, Benefit of items in SlaverSite1

B	Tidset	2	6	7	9	10
	Benefit	60	10	100	10	60
C	Tidset	1	3	5	8	10
	Benefit	16	1	4	25	2

Table 8. Local TWU of SlaverSite1

TID	Local TWU
T ₁	16
T ₂	60
T ₃	7
T ₄	3
T ₅	4
T ₆	13
T ₇	100
T ₈	34
T ₉	13
T ₁₀	62

SlaverSite1 does not send information of item A to MasterSite because its *twu* value does not satisfy *minutil*. It helps the algorithm saving time for communication between SlaverSites and MasterSite, and saves memory in MasterSite.

SlaverSite2: **Table 9.** Tidset, Benefit of items in SlaverSite2

D	Tidset	2	4	7	8				
	Benefit	6	6	6	18				
E	Tidset	1	2	3	4	5	7	8	10
	Benefit	5	5	5	5	10	5	5	10

Table 10. Local TWU of SlaverSite2

TID	Local TWU
T ₁	5
T ₂	11
T ₃	5
T ₄	11
T ₅	10
T ₇	11
T ₈	23
T ₁₀	10

SlaverSite1 and SlaverSite2 send their local TWU to MasterSite, and MasterSite computes global TWU by adding all *twu* values which are the same *tid*.

Then, MasterSite sends global *twu* values to all SlaverSites and receives information from SlaverSites as in Table 7 and Table 9. When MasterSite receives enough information, the DTWU-Mining algorithm will be called.

4 Experiments

Algorithm was coded by C# 2005. Configuration of PC using for experiments evaluating includes CPU Intel 2.0 GHz Centrino, RAM 1 GB, Windows XP. Experimental databases have features such as follow:

Table 11. Experimental databases

Database	#Trans	#Items	Remark
BMS-POS	515597	1656	Modified
Retails	88162	16469	Modified

We modified by adding one more column contains values (random from 1 to 10) for each item corresponding to each transaction, and create one more table to store benefit values of items (value from 1 to 10). Each database is distributed into 5 Sites.

Because TWU-Mining [6] is often faster than algorithms based on utility upper bound [15, 16] and Two-Phase [7], we only compare proposed algorithm with TWU-Mining.

Table 12. Experimental results

databases	<i>minutil</i> (%)	TWU-Mining (s)	Proposed alg. (s)	#HUIs
BMS-POS	5	27.59	18.54	4
	4	39.05	23.60	6
	3	55.67	31.09	7
	2	95.56	49.42	22
Retails	0.8	11.31	7.36	29
	0.6	23.23	12.68	45
	0.4	57.69	28.96	64
	0.2	178.19	89.25	239

Table 12 shows that the mining time of parallel algorithm is less than that of TWU-Mining on centralized database [6].

5 Conclusion and Future Works

This paper has proposed the parallel method for mining HUIs from vertically partitioned distributed databases, and the efficient algorithm is also proposed. As above mentioned, the mining algorithm in distributed databases is more efficient than that in centralized database [6]. By WIT-tree technique, the algorithm only scans local databases one time and only item that its *twu* satisfies *minutil* must be sent to MasterSite. Therefore, it spends a little time for communication between MasterSite and SlaverSites.

In the future, we are also interesting in developing of HUIs applications such as mining association rules from HUIs.

In this paper, we only study the method for mining HUIs from vertically partitioned distributed databases. An efficient algorithm for mining HUIs in horizontally partitioned distributed databases will be discussed in the future. Besides, grid computing for mining HUIs in distributed databases will be discussed.

References

1. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules. In: VLDB 1994, pp. 487–499 (1994)
2. Agrawal, R., Shafer, J.: Parallel Mining of Association Rules. IEEE Trans. Knowledge and Data Eng. 8(6), 962–969 (1996)

3. Erwin, A., Gopalan, R.P., Achuthan, N.R.: CTU-Mine: An Efficient High Utility Itemset Mining Algorithm Using the Pattern Growth Approach. In: IEEE 7th International Conferences on Computer and Information Technology, Aizu Wakamatsu, Japan, pp. 71–76 (2007)
4. Erwin, A., Gopalan, R.P., Achuthan, N.R.: A Bottom-Up Projection Based Algorithm for Mining High Utility Itemsets. In: The 2nd International Workshop on Integrating Artificial Intelligence and Data Mining, Gold Coast, Australia, vol. 84, pp. 3–11 (2007)
5. Khan, M.S., Mueyba, M., Coenen, F.: A Weighted Utility Framework for Mining Association Rule. In: Proc. IEEE European Modelling Symposium (EMS 2008), pp. 87–92 (2008)
6. Le, B., Nguyen, H., Cao, T.A., Vo, B.: A Novel Algorithm for Mining High Utility Itemsets. In: IEEE 1st Asian Conference on Intelligent Information and Database Systems, Quang Binh, Vietnam, April 1-3, pp. 13–17 (2009)
7. Liu, Y., Liao, W., Choudhary, A.: A Fast High Utility Itemsets Mining Algorithm. In: UBDM 2005, Chicago, Illinois, USA, August 21, pp. 90–99 (2005)
8. Luo, P., Xuong, H., Lu, K., Shi, Z.: Distributed Classification in Peer-to-Peer Networks. In: KDD 2007, San Jose, California, USA, August 12–15 (2007)
9. Miller, D.J., Zhang, Y., Kesidis, G.: Decision Aggregation in Distributed Classification by a Transductive Extension of Maximum Entropy/Improved Iterative Scaling. EURASIP Journal on Advances in Signal Processing 2008 (2008)
10. Schuster, A., Wolff, R.: Communication-efficient distributed mining of association rules. In: Proc. of the 2001 ACM SIGMOD Int'l. Conference on Management of Data, Santa Barbara, California, pp. 473–484 (2001)
11. Serazi, M., Perera, A., Abidin, T., Hamer, G., Perrizo, W.: An API for Transparent Distributed Vertical Data Mining. In: Proceedings of the ISCA 14th International Conference on Intelligent and Adaptive Systems and Software Engineering, Novotel Toronto Centre, Toronto, Canada, July 20-22, pp. 151–156 (2005)
12. Suchahyo, Y.G., Gopalan, R.P.: CT-PRO: A Bottom-Up Non Recursive Frequent Itemset Mining Algorithm Using Compressed FP-Tree Data Structure. In: IEEE ICDM Workshop on Frequent Itemset Mining Implementation (FIMI), Brighton UK (2004)
13. Wolff, R., Schuster, A.: Association Rule Mining in Peer-to-Peer Systems. IEEE Trans. Systems, Man and Cybernetics, Part B 34(6), 2426–2438 (2004)
14. Yao, H., Hamilton, H.J., Butz, C.J.: A Foundational Approach to Mining Itemset Utilities from Databases. In: Proceedings of 2004 SIAM International Conference on Data Mining, pp. 482–486 (2004)
15. Yao, H., Hamilton, H.J.: Mining Itemset Utilities from Transaction Databases. Data and Knowledge Engineering 59, 603–626 (2005)
16. Yao, H., Hamilton, H.J., Geng, L.: A Unified Framework for Utility Based Measures for Mining Itemsets. In: UBDM 2006, Philadelphia, Pennsylvania, USA, pp. 28–37 (2006)
17. Yu, G., Shao, S., Sun, D., Luo, B.: Mining Long High Utility Itemsets in Transaction Databases. WSEAS Transactions on Information Science & Applications 5(2), 326–331 (2008)
18. Zaki, M.J., Hsiao, C.J.: Efficient Algorithms for Mining Closed Itemsets and Their Lattice Structure. IEEE Transactions on Knowledge and Data Engineering 17(4), 462–478 (2005)
19. Zaki, M.J.: Parallel and Distributed Association Mining: A Survey. IEEE Concurrency, Special Issue on Parallel Mechanisms for Data Mining, 14–25 (December 1999)

An Ontology-Based Autonomic System for Improving Data Warehouse Performances

Vlad Nicolicin-Georgescu^{1,2}, Vincent Benatier², Remi Lehn¹, and Henri Briand¹

¹LINA CNRS 6241 - COD Team - Polytech'Nantes,
Site Ecole Polytechnique de l'université de Nantes, Rue Christian Pauc ,
44306 Nantes, France

henri.briand@univ-nantes.fr

²SP2 Solutions, 8 Rue Rene Coty, 85000 La Roche sur Yon, France

vladgeorgescun@sp2.fr

www.sp2.fr

Abstract. With the increase in the amount and complexity of information, data warehouse performance has become a constant issue, especially for decision support systems. As decisional experts are faced with the management of more complex data warehouses, a need for autonomic management capabilities is shown to help them in their work. Implementing autonomic managers over knowledge bases to manage them is a solution that we find more and more used in business intelligence environments. What we propose, as decisional system experts, is an autonomic system for analyzing and improving data warehouse cache memory allocations in a client environment. The system formalizes aspects of the knowledge involved in the process of decision making (from system hardware specifications to practices describing cache allocation) into the same knowledge base in the form of ontologies, analyzes the current performance level (such as query average response time values) and proposes new cache allocation values so that better performance is obtained.

Keywords: Business Intelligence, Decision Support Systems, Autonomic Computing, Data Warehouse, Ontology, Cache, Business Rule.

1 Introduction

As the 21st century is well on its way, in a civilized and modern world, we realize that the most important asset needed in order to keep the pace with this new rhythm is knowledge. Knowledge is the source of power and truly the new edge of the power shift [4]. As technology comes greatly to our help, we find it normal to research, discover and improve ways of gathering, processing and using all information available. Our purpose is to develop a system aimed at helping enterprises analyze and improve their decision making process by providing a unified representation of certain aspects involving the knowledge available for this process (from software support documents to human expert experience) and an autonomic system that makes use of this knowledge and acts upon it. Simpler and often referred to as DSSs, *Decision Support Systems* are defined as computerized systems whose main goal is to analyze a series of facts and give various propositions for actions regarding the facts involved [12]. This

is why the process of decision making, based on such systems and the elements involved, is known as *business intelligence* (BI) process.

The applicative area of this paper is *cache memory* allocation for the Oracle Hyperion Essbase BI¹ cubes. This is a common configuration problem that BI experts are faced with. The cubes represent the data warehouse whose performances are to be improved. The system we propose makes use of knowledge based on system information (from architecture to cube cache parameters) and on sets of rules representing constraints and advice for the cache allocations (taken from the Essbase documents and from our human experts). The purpose is to provide two main functionalities. First, to compute a system's degree of improvement based on cache allocations and performance indicators. Second, to propose an improved cache configuration, that gives (if possible) the optimal performances. Two main aspects have been taken into consideration along with this approach.

The first aspect is knowledge representation. The knowledge regroups several sources: describing software and hardware architectures, system performance measurement, system analysis and improvement practices (described as sets of ECA (event condition action) rules [11]). Knowledge representation describes how this information is unified into knowledge bases. If for the system architecture, models are being developed and even adopted as w3c standards², then for the data representation of system report performances and the rules of system analysis, we are obliged to turn to specific representations (using ontologies [16] and ontology based rules).

Second, the improvement process itself, meaning having a fast response (from the moment a demand for improvement is made) and having a good (if not the best) response for any type of decision request (in our case a new cache allocation). In order to achieve this, IBM has proposed a solution to help automate various processes. The solution is called *Autonomic Computing* [6], [11] and its applications extend way beyond the business intelligence sector. We propose the usage of autonomic computing with the cache allocation improvement process.

Section 2 gives an insight of how we manage the knowledge in our system in order to drive the data warehouse. First we present how the knowledge base is organized for managing data warehouses and then how autonomic computing is used in the decision making process. Section 3 focuses on the description of our model and how this approach is used to perform analysis and improvement. A schema of a DSS together with the description of the data warehouses and an example of associated rules are presented. Section 4 provides a view of the experimentation and the results obtained. Finally, we sum up the work presented and take a glance at the future directions.

2 Data Warehouse Management

2.1 Knowledge Management

In brief, Knowledge Management is the process through which organizations generate value from their intellectual and knowledge-based assets, disseminating this knowledge

¹ http://download.oracle.com/docs/cd/E10530_01/doc/epm.931/html_esb_dbag/frameset.htm?dstcache.htm

² <http://www.w3.org/TR/2008/CR-sml-20081125/>

and sharing it in an effort to get competitive advantage [7]. Data warehouse (DW) management is a key element in the decision making process. A data warehouse is a repository of an organization's electronically stored data and is designed to facilitate reporting and analysis [20]. Managing a data warehouse includes the process of analyzing, extracting, transforming and loading data and metadata. Our interest in knowledge management comes from the types of data involved in the decision making process.

The knowledge management into our work is based on system analysis and functioning. These refer to a complex set of rules that describes the functioning and non trivial interdependencies between the elements of the system. The main objective is to describe the rules for the analysis and improvement processes. Representing data under the form of rules [14] gives a completely different approach to knowledge management. Practically, we create a business rule knowledge base that serves for the process of analysis and improvement. This process is supported both by the human expert and by the autonomic system. We propose to divide these rules into two main components: constraints and advice.

Advice represents business rules (BR) and best practices for the DSS giving the measure of a system improvement level in these terms. This means how 'close' a configuration is to satisfy sets of advice and therefore is able to generate an advice scoring. This scoring is built upon a point allocation system that grades the level of implementation of an advice set which we call a BR improvement points system, and which we describe in Section 3.

Constraints represent limitations imposed (i.e. the index cache cannot be under 1 Mb) and a violation of such constraints leads to an error in the system analysis.

To the division above, an entire set of rules is added, (from initial fact deduction to planning and action rules). These sets of rules are considered as state specific rules and are modeled for each state of the autonomic computing manager (presented in the next section).

2.2 Autonomic Computing in Data Warehouse Management

Most of the IT organizations spend a lot of time reacting to problems that occur at the IT infrastructure component level. This prevents them from focusing on monitoring their systems and from being able to predict and prevent problems before end users are impacted [5]. Autonomic computing (AC) is the ability for an IT infrastructure to adapt and change in accordance with business policies and objectives. Quite simply, it is about freeing IT professionals to focus on higher-value tasks by making technology work smarter, with business rules guiding systems to be self-configuring, self-healing, self-optimizing and self-protecting [6]. This subject is of great interest to enterprises and has already been put into practice for improving database performance by IBM [19], [15] and Microsoft [2]. There is great interest of development into applications of autonomic computing on managing data warehouses, as experts can no longer face the quantity of information available.

IBM specifications link autonomic computing with the notion of *autonomic manager* as the entity that coordinates the activity of the autonomic process. Four separate phases are distinguished for the manager: monitoring, analyzing, planning and executing [6], [10]. We propose an implementation of the autonomic manager connected to our knowledge base and based on the data warehouse performance. As it is rule based

knowledge, we differentiate the sets of rules for each of these phases. The illustration of this process is shown in Section 3. Similar alternatives to autonomic computing were made in real BI [18] but the idea is the same: to be able to analyze and improve (in our case) a given system through a closed loop that differentiates a series of states.

The loop formed by the four states mentioned is regularly run. By regularly we mean once per night during batch operations, when statistics on the data warehouse usage for that day are gathered. Each loop, according to the new query times, modifies the values of the caches, and this is repeated until the desired times are achieved. This process is based on both the feedback from the previous response times and on the sets of advice mentioned earlier. Consequently, what our solution proposes is to include business rule in the loop so that the modifications to the cache values are more substantial and relevant, and thus the time needed to reach the desired performance level is greatly reduced.

3 Knowledge Base

We have seen the information we need to formalize and we have found the ontology [16] as a model of knowledge representation. The ontology representation suits our needs as it provides the solution for two main problems: knowledge unification and knowledge interchange. Works in this area have already been done by [9] and we found this model fully applicable to our system as it covers both knowledge formalization and rule usage. We propose a division of the knowledge aspect into two main categories: static and dynamic.

3.1 Static Knowledge Base

The static aspect of knowledge contains all the knowledge representation under the form of ontology concepts: classes, individuals and the properties linking them. Our implementation uses OWL³ as ontology description language and Protégé⁴ as software support for ontology manipulation. The ontology contains over 150 concepts and over 250 axioms. We propose two main data types for the static knowledge base:

System information and architecture refers to all data concerning the software and hardware specifications of the DSS system (i.e. the quantity of RAM memory installed on a server). This subject has been approached [1] and detailed by a certain number of editors [13]. Fig. 1 shows how we model the DSS architecture hierarchically, with the use of a UML⁵ diagram. Several entities are distinguished, starting from the top of the hierarchical tree (the DSS) and, at each level one or more sub-components can be identified with the specific parameters.

System report and performance contains aspects regarding the DSS, in particular the data warehouse, parameters and performance indicators. There are many indicators to take into consideration: memory cache values, query response times, report editing times, aggregation operation times, etc. This paper considers three types of the

³ http://www.w3.org/2007/OWL/wiki/OWL_Working_Group

⁴ <http://protege.stanford.edu/download/registered.html>

⁵ <http://www.uml.org/>

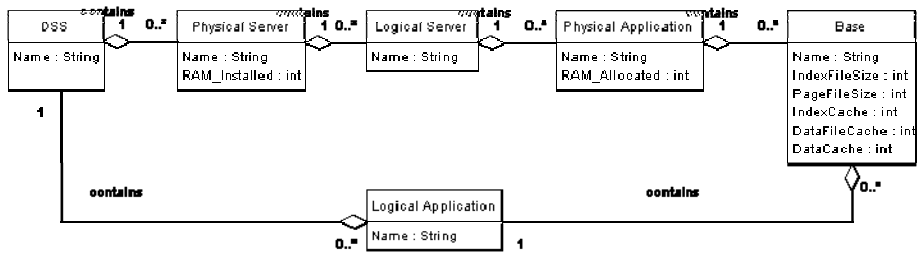


Fig. 1. DSS hierarchical architecture

Essbase cube cache: Index Cache, Data File Cache and Data Cache and the query response time as performance indicator. The implication is that by a good configuration of these three caches according to a given situation (i.e. at night some bases are not used, some cache allocations are useless, etc.) we gain substantial processing time. Although the problematic of performance improvement in data warehouses throughout caches is debated [17], [3] the issue is always addressed either through the physical design or the design of algorithms to determine which information is likely to be stored in cache memories. Cache allocation improvement is an important aspect of data warehouse tuning and there are little or no works on cache improvement in the context of data warehouses.

3.2 Dynamic Knowledge Base

The dynamic aspect is the challenge in our work and provides the main innovating approaches. It formalizes the functioning system and the analysis aspect of data warehouse management presented in 2.1. It contains all the rules that are part of the AC loop and all the rules that determine property interdependence and individual inference in the ontology. For reasons of simplicity and efficiency we have chosen to use Jena Rules via the Jena Java API⁶ for ontology development. The rules are divided according to their area of activity in conformity with the AC loop phases.

For the business rule illustration, we give below an example of an analysis business rule that informally states: *The closer the Index Cache value is to the Index File Size for a base, the better.* We formalize this rule and obtain a discrete point allocation for the different proportions:

```

[rule1: (?base rdf:type cp:c_Base) (?base cp:dp_hasIndexFileSize ?ifs) (?base
cp:dp_hasIndexCache ?ic) quotient(?ic, ?ifs, ?rap) ge(?rap, "0.95"^^xsd:double)
-> (?base cp:dp_hasPoints_Advice_IndexCacheAllocation "1000"^^xsd:int)]
  
```

```

[rule2: (?base rdf:type cp:c_Base) (?base cp:dp_hasIndexFileSize ?ifs) (?base
cp:dp_hasIndexCache ?ic) quotient(?ic, ?ifs, ?rap) lessThan(?rap, "0.95"^^xsd:double)
ge(?rap, "0.85"^^xsd:double)
-> (?base cp:dp_hasPoints_Advice_IndexCacheAllocation "900"^^xsd:int)] ...
  
```

⁶ <http://jena.sourceforge.net/inference/#rules>

The formalization process of business rules such as this one a very important aspect, and always requires an expert hand. By applying all the analysis rules we obtain an overall scoring and we can calculate a performance level of all the data warehouses from the point of view of business rules.

For the autonomic computing rule illustration we propose a simple example of the passage through the 4 states of the autonomic manager. We show the index cache is modified in a cycle.

Monitor: retrieval of the response time and the current cache values for a DW. These are stored in the knowledge base via the java program.

Analyse: we compare the average response time of the DW with its desired response time. If it is greater, the caches must be increased:

```
(?base cp:dp_hasAvgResponseTime ?avgt) (?base cp:dp_hasDesiredResponseTime ?dt)
ge(?avgt, ?dt) -> (?base cp:hasState cp:IncreaseCache)
```

Plan: we try to see if the increased cache state can be applied for the index cache. If the new cache value (increased with 10% of its current value) is not greater than the allocated memory, then plan the change to the index cache:

```
(?base cp:hasState cp:IncreaseCache) (?base cp:dp_hasIndexCache ?ic) (?base
cp:dp_hasAllocatedMemory ?am) product(?ic, '1.1'^xsd;double, ?newic) le(?newic, ?am) ->
(?base cpdp_:dp_hasIndexCache ?newic)
```

Execute: Execute a modification script with the new proposed value of the index cache. This is done via the java program.

4 Experimentation and Results

For our experiments we have considered a test suite that simulates a real environment with the associated parameters. On an existing server we have chosen an Essbase cube as the data warehouse whose performances were improved. For the pertinence of the tests, the cube was created starting from the “Sample” base provided by Essbase. The cube contains in average 11 principal axes and 27 level 2 axes and the data file has an average size of 300MB.

With the respective cube we carried out several tests corresponding to several configurations. We had to simulate the night/day loop passage faster so we made a series of 5 queries (from very fast to very slow as time of response) and applied them to each configurations. This process was iterated 10 times therefore simulating a day/night cycle. At the end of each cycle we fetch the average response times for each of the bases and pass through the autonomic computing loop so that we could optimize the cache allocation where it was necessary. The evolution of cache allocation with the response times can be observed in Fig. 2. We can see the difference between the minimum cache allocations in configuration C1 which is almost 6 times slower than the maximum possible allocation in configuration C6. As a maximum allocation is not always possible due to the quantity of memory available, we have to try our best to improve the performances. Applying our approach, configuration C3 maximizes the BR improvement points, and with a passage through the AC loop several

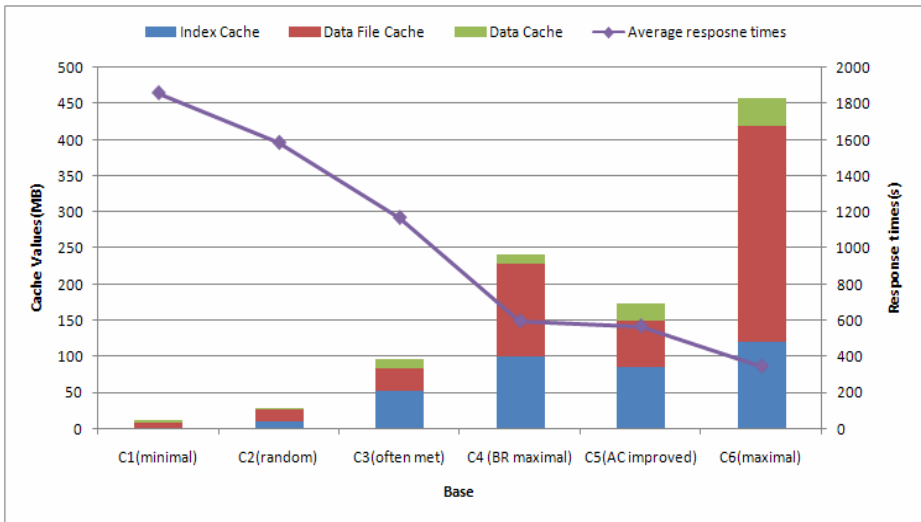


Fig. 2. Query average response times evolution with the cache configurations

times we get a configuration which has a smaller response time C4. The main improvement is that C4 is using only 174MB for its caches whereas C3 requires 240MB. In addition, the average response times in comparison with a normal or random configuration (as in C2 and C3) are improved at least 2 times.

5 Conclusions

This article presented how autonomic computing and ontologies can be used for helping DSS experts improve the cache memory allocations for data warehouses. It is not the first approach that tries to combine the two elements together [9],[8] but the premiere is its application in the field of business intelligence and data warehouse improvement using business rules.

There are many positive aspects to this approach of decision support system management: the simple and intuitive yet powerful ontology representation, the facilities of web semantics and rule support brought to the domain of BI and, last but not least, the process of autonomic computing to manage all these elements.

Our future directions are to expand the Data Warehouses described above so that our small prototype can prove its efficiency on more than a 'few simple rules'. Our purpose is to integrate the prototype presented here with more than one aspect (data warehouse cache allocations based on response times) of decision support systems.

As the domain is relatively new and not much has been written on the subject yet, we try to bring as much support as possible for future development in the direction of autonomic decision support systems. We follow these changes and hope that our work will equally add something to this expanding environment.

References

1. Ackerman, A., Tyree, J.: Using ontologies to support development of software architectures. *IBM Systems Journal* 45(4) (2006)
2. Mateen, A., Raza, B., Hussain, T.: Autonomic Computing in SQL Server. In: 7th IEEE/ACIS International Conference on Computer and Information Science, pp. 113–118 (2008)
3. Saharia, A.N., Babad, Y.M.: Enhancing Data Warehouse Performance through Query Caching. *The DATA BASE Advances in Informatics Systems* 31(3) (2000)
4. Toffler, A.: *Powershift: Knowledge, Wealth and Power at the edge of the 21st Century*. Bantam Book Publishing (1991)
5. Manoel, E., Nielsen, M.J., Salahshour, A., Sampath, S., Sudarshanan, S.: Problem determination using self-managing autonomic technology. *IBM RedBook*, 5–9 (2005)
6. IBM Corporation, An architectural blueprint for autonomic computing, pp. 9–18, et Autonomic Computing. Powering your business for success. *International Journal of Computer Science and Network Security* 7(10), 2–4 (2005)
7. Oliveira, J., de Souza, J.M., Perazol, M.: Managing knowledge about resources for autonomic computing. In: 1st latin american autonomic computing symposium, pp. 124–126 (2006)
8. Gonzales, J.M., Lozano, J.A., Lopez de Vergara, J.E., Villagra, V.A.: Self-adapted service offering for residential environments. In: *ACNM 2007*, pp. 48–55 (2007)
9. Stojanovic, L., Schneider, J., Maedche, A., Libischer, S., Studer, R., Lumpp, T., Abecker, A., Breiter, G., Dinger, J.: The role of ontologies in autonomic computing systems. *IBM Systems Journal* 43(3), 598–616 (2004)
10. Parshar, M., Hariri, S.: *Autonomic Computing: Concepts, Infrastructure and Applications*. Taylor and Francis Group, Abington (2007)
11. Huebscher, M.C., McCann, J.A.: A Survey on Autonomic Computing – Degrees, Models and Applications. *ACM Computing Surveys* 40(3), Article 7 (2008)
12. Druzdal, M.J., Flynn, R.R.: *Decision Support Systems*. Encyclopedia of library and information science (1999)
13. Microsoft Corporation, Understanding system definition model (SDM) and its practical application in 2006 to 2008, pp. 3–5 (2006)
14. Stojanovic, N., Handschuh, S.: A framework for knowledge management on the semantic web. In: *The 11th International WWW Conference* (2002)
15. Lightstone, S.S., Lohman, G., Zilio, D.: Toward autonomic computing with DB2 universal database. *ACM SIGMOD Record* 31(3) (2002)
16. Gruber, T.: *What is an ontology?* Academic Press Pub., London (1992)
17. Malik, T., Wang, X., Burns, R., Dash, D., Ailamaki, A.: Automated Physical Design in Database Caching. In: *ICDE Workshop* (2008)
18. Nguyen, T.M., Schiefer, J., Min Tjoa, A.: Sense & Response Service Architecture (SARESA). In: *DOLAP 2005* (2005)
19. Markl, V., Lohman, G.M., Raman, V.: LEO: An Autonomic Optimizer for DB2. *IBM Systems Journal* 42(1) (2003)
20. Inmon, W.H.: Tech topic: what is a data warehouse? *Prism solutions* 1 (1995); *Data warehouse performance*, pp. 19–20, 209–304. Wiley Publishing, Chichester (1999) and *Building the data warehouse*, 4th edn., pp. 29–33, 79–94, 331–333. Wiley Publishing, Chichester (2005)

Semantic Enhancement of the Course Curriculum Design Process

Javier Vaquero¹, Carlos Toro¹, Juantxu Martín², and Andoni Aregita³

¹ VICOMTech Research Centre, Mikeletegi Pasalekua 57, 20009 San Sebastian, Spain

² BIB S. Coop, Eibar, Spain

³ Alecop, S. Coop., Arrasate, Spain

Abstract. In this paper we propose a methodology intended to improve Course Curriculum Design (CCD) tasks, using for this purpose Semantics and CBR techniques. In specific, our proposed methodology is focused on two points: (i) the re-use of available resources (courses, etc), and (ii) the application of the experience of different experts in the course creation. As a prove of concept, we present a case study where our methodology is applied for competence and course creation using the Spanish normative for vocational education domain (technical degree).

Keywords: Ontologies, Case-Based Reasoning, Course Design, Knowledge Based Systems.

1 Introduction

Education quality is related to the best use of the available resources, the proper design of the subjects and evaluations, and generally, to a good design of the courses and components which are part of the education process [1]. Typically, course design starts with the definitions of the competences or abilities that must be met at the end of the course. As pointed by Diamond [2], educators need to clearly identify goals prior to any kind of course assessment. Those goals are the same to what we understand as competence evaluation. Based on competences, the course designer builds the contents and the evaluations in a process known as Course Curriculum Design (CCD) [2]. We have found that CCD presents some interesting challenges. From the computational perspective, some of the most interesting are the following:

- Every country has its own design normative, and successful experiences in one country cannot be easily applied to another.
- Designers have different points of view, and the same course design differs from one designer to other. This situation leads to non-homogenized curriculum.
- The re-use of knowledge and prior user experiences coming from different experts is not included in the approach

In this paper, we propose a novel approach for the Course Curriculum Design that is enhanced by semantics, taking into account course designers' expertise and case based reasoning systems in order to produce a better course with the available resources. This

paper is structured as follows: in section 2, we introduce a brief explanation about related concepts. In section 3, we present our proposed methodology. In section 4, we present a case study where proposed methodology is implemented. And finally, in section 5, we present our conclusions and future work.

2 Related Concepts

In this section, we introduce some concepts relevant to this paper. Our intention is to give a short overview of the involved technologies. An interested reader is invited to review [3], [4], [5] for a wider explanation on the concepts presented.

2.1 Case Based Reasoning (CBR)

CBR is a problem solving technique based on two tenets: *(i)* the world is regular, so many similar problems have similar solutions, and *(ii)* types of problems an agent encounters tend to recur [6]. CBR does not use generalized rules as a knowledge source, but a memory of stored cases recording specific prior episodes [6]. New solutions are generated by retrieving the most relevant cases from memory and adapting them to fit new situations. This is a powerful and frequently used way to resolve problems by humans. In its simplest form, CBR has four steps: *(i)* situation assessment, *(ii)* case retrieval, *(iii)* similarity evaluation and *(iv)* storage of the new case [6]. CBR as a tool is an important aid to analyze previous decisions by using statistical models. One of the keys success factors in using CBR based tool is the fact that every new choice made upon the available cases, feed back to the database and therefore enhancing the model. We believe that by mixing CBR and Semantic technologies the strong points of both techniques can be leveraged to the users advantage, while at the same time their weakness can be alleviated.

2.2 Semantic Technologies

Semantics is the area of the knowledge that studies the meaning of things [7]. Semantic technologies constitute one of the more interesting technologies derived from the World Wide Web revolution. In this work, we use ontology modelling for its inference capabilities and to support our architecture from a knowledge engineering point of view. Next section introduces brief descriptions regarding the semantic based technologies relevant to our work.

2.2.1 Ontologies

There are many possible definitions to describe what ontology is. In the Computer Science domain, the widely accepted definition states that “an ontology, is the explicit specification of a conceptualization” [8], or in other words an ontology is the description of the concepts and relationships in a domain of study. Some of the motivations to model a domain with ontologies are *(i)* to share common understanding of the structure of information among people or software agents, *(ii)* to enable reuse of domain knowledge, to make domain assumptions explicit, *(iii)* to separate domain knowledge from the operational knowledge, and *(iv)* to analyze the domain’s modelled knowledge. Ontologies can be modelled using different languages, for instance,

RDF, RDFS and OWL, the later is a new standard from the W3C consortium available in three flavours, OWL-Lite, OWL-DL and OWL-Full, depending on the desired level of semantic load [9]. The main characteristic of an ontology-based solution is its capacity to semantically infer newly derived information. Such information is not explicitly specified by the user and in order to obtain it modern inference engines and reasoners, like Racer or Pellet [10], are used.

2.3 The Set of Experience Knowledge Structure (SOEKS)

The Set of Experience Knowledge Structure is an experience tool able to collect and manage explicit knowledge of different forms of formal decision events [5]. The SOEKS has been developed as part of a platform for transforming information into knowledge named Knowledge Supply Chain System. In this methodology, there are four basic components: variables, functions, constraints and rules associated and stored in a combined dynamic structure. The combination of the four components of the SOEKS offers distinctiveness due to the elements of the structure that are connected among themselves, imitating part of a long strand of DNA. A SOEKS produces a value of decision, called efficiency. Besides, it is possible to group sets of experience by category, that is, by objectives. These groups could be store a “strategy” for such category of decisions. By using this methodology and its underlying ability to model user experience, we have a proven method for modelling the Expert who is designing the Course. The benefits of using such expert model is not only restricted to the fact that the actual system becomes more stable in terms of repeatability (the same set of needs will produce similar courses) but in terms of confidence and trust as the Case based systems produce only statistical results on the set of variables, while by using expert models, such systems would be filtered to enhance in fact the CCD.

3 Proposed Methodology

In this section, we propose our methodology for the enhancement of the CCD process. The goal of our methodology is to solve two common reported problems: (i) CCD uses many different sources of information (in many cases, in non-digitalized forms), and (ii) At present time, a CCD expert is not able to use the prior experiences from other CCD experts in the design of new similar courses. In Fig. 1, our methodology is depicted. As can be seen, we divided it in four stages that will be explained next:

3.1 Data Categorization

On a normal scenario, initial information for the CCD is found dispersed in different media types (e.g. electronic, books, similar courses and their components, etc). The aforementioned data needs to be categorized in usable differentiated types of elements for posterior analysis. In the first stage of our methodology, a human practised sorts out the information using their knowledge of the domain; possible categorizations are made using domain restrictions and such domain depends on the specific educational normative used. The need for an expert in this stage obeys to the necessity of comprehension on the characteristics of the courses that is a direct consequence of the

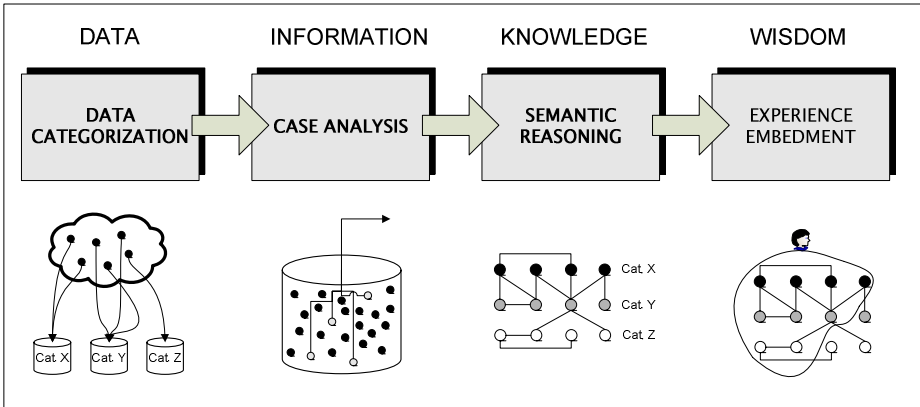


Fig. 1. The four stages of our proposed methodology

knowledge about the domain. Usually, in this stage some computational techniques could be used, e.g. Data Mining, taxonomies, however the most common approach is a manual sort. At the time being, this part of our methodology is the only one that needs direct human interaction, although we are currently pursuing some lines of research that could lead us to automate it. Typical data categorization for an educational domain could be for example: courses, competences, objectives, evaluations, etc.

3.2 Case Analysis

Once the data is organized, the contained elements are ready to be queried in order to solve the question of which ones better answer the set of requirements for the course. In order to perform the aforementioned information query, relational databases and Case Based Reasoning are two techniques that can be used (this last technique was our choice for implementation due to easiness on implementation). A typical query for a Case analysis could be: *“retrieve the available courses of less than 200 hours with algebra pre-requisites”*. The goal of this stage is to obtain full advantage of the stored data.

3.3 Semantic Reasoning

Until this point, the elements belonging to the different categories can be recovered in order to compose an answer to a complex question. For such query retrieval, it is necessary to know how these elements are related. As the domain defines relationships, so there exists the need of a model of such specific domain. This model can be obtained using different Domain Modelling techniques, and stored in several ways, like relational databases or ontologies. When ontologies are the election, there is possible to infer semantically new information not stored explicitly, using semantic reasoners. A typical question could be: *“Retrieve the competences that have the needing of an advanced algebra course”*.

3.4 Experience Embedment

Lastly, the expert experience has to be used in order to refine the obtained answer. As mentioned before, experts have different perspectives of how the same course has to be assembled, for example, an expert in CCD with an engineering background, would design a physics course oriented to engineering better than a physicist would do. In this last stage of our methodology, the goal is to apply the experience of experts with the correct point of view for the specific situation. For this purpose, a model that allows storing and handling of such experience is needed. SOEKS is an experience modelling and storing structure that we used in several knowledge systems previously [11], obtaining good results.

4 Case Study

In this section, we present our case study; we intend to model a new course using our methodology with compliance to the Spanish normative for vocational education domain (technical degree). With the help of an expert in this domain, we identified five different categories to distribute the data: (i) competences, i.e. the abilities to be acquired, (ii) objectives, i.e. the necessary goals to achieve the target competence, (iii) courses, i.e. individual and self-contained pieces of knowledge that fulfills an objective, (iv) units and (v) projects, being both of them the smallest elements belonging to courses. For each one of the five categories, we created a repository to store the elements belonging using for such purpose a MySQL database. We have found that separated repositories improve the extensibility, modularity and efficiency of the system. Upon the repositories, we implemented CBR modules for the Case Analysis

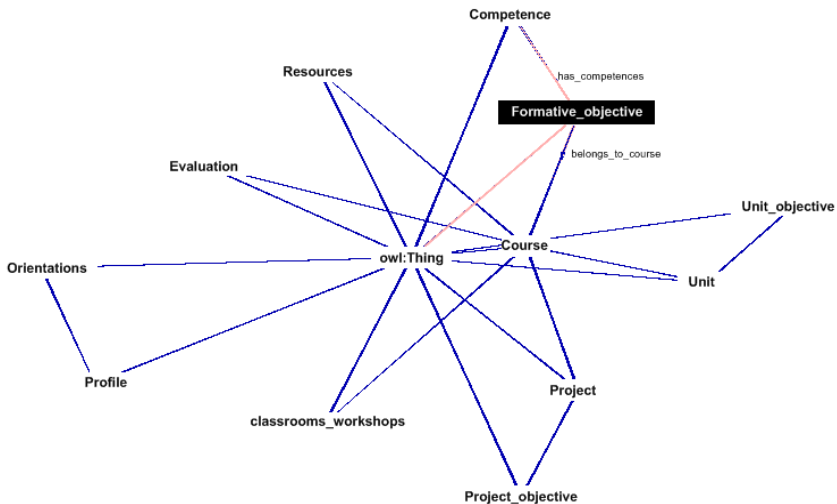


Fig. 2. Graphical representation of part of the set of ontologies

stage. Each category (except objectives) has its own independent CBR module, in order to maintain the modularity. These CBR modules have been developed with jColibri2 [12], a Java API for the development of CBR systems created by the GAIA group at the Complutense University of Madrid. For the implementation of the Semantic Reasoning stage, we modelled the domain in a set of ontologies in OWL-DL [9]. The use of OWL-DL was decisive when reasoning the Knowledge Base from the open world assumption as it simplified the rule system used (not shown here for confidentiality reasons with the contractor of the underlying project which generated this approach). The modelling process was performed in the Protégé ontology editor, in part because of the possibility of use of its API for the automatic instantiation. A section of the resultant ontology is depicted in Fig. 2.

Queries are made to the ontology using the Protégé OWL API [13] and reasoned using Pellet [10] in order to infer new knowledge. The last stage of the methodology comprises the experience filtering by using the model of the expert with a SOEKS compliant methodology [14]. Fig. 3 depicts the scenario described before.

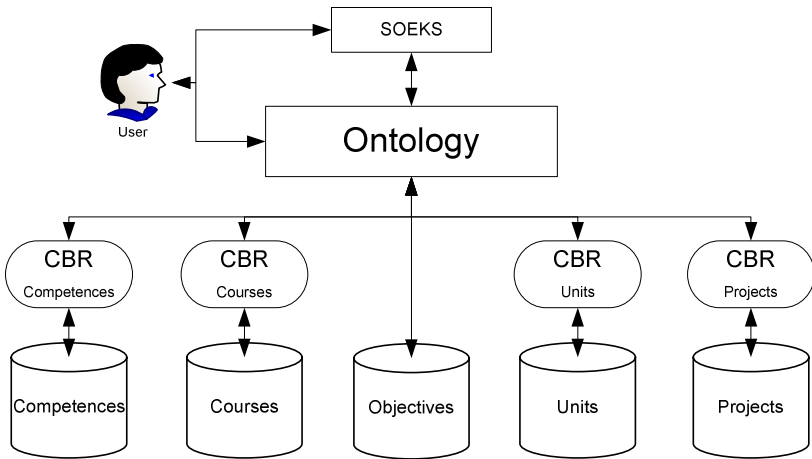


Fig. 3. Implementation for our case study

Fig. 4 depicts a user case for competence and course creation, divided in three different tasks: competence creation, course creation, and objective assignment.

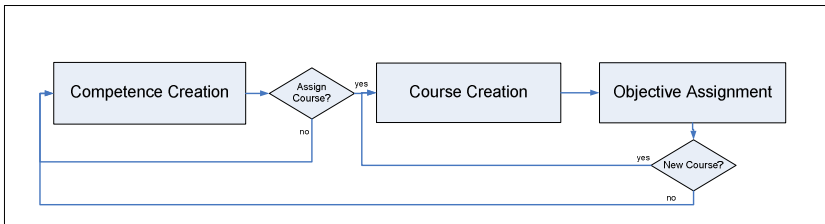


Fig. 4. Create courses for a non-existing competence user case

For the competence creation, the user defines the master guides that their new competence has to fulfil. With these guides, the CBR module implemented over the competences repository is launched, and returns a set of competences that are near to the desired new competence. With the help of these returned competences, the user is able to complete their new competence. CBR can be launched more times if necessary, changing the master guides in order to obtain new suggested competences. Once is completed, the new competence is stored and becomes part of the stored cases of the CBR for future questions. Once the new competence creation is done, the new task is the creation of courses for such competence. Using the relations between elements stored in the ontology, the system offers to the user the courses (and the units and the projects that form them) related to the competences used for the new competence creation. Moreover, user can decide to launch CBR processes over courses, units or projects, in order to obtain more suggestions for their design of the course. When a course is completed, user has to assign a common objective to the competence and to the course. As we said before, in our domain objectives are quite simple elements, so we decide not implement the option to launch any CBR process over the repository of objectives but show all them in a list. If desired objective is in the list, the user chooses it. If the objective is not in the list, the user creates it directly. When this task ends, the course is stored in its repository and the user can create another course for the competence or end the user case.

5 Conclusions and Future Work

In this work, we presented a new methodology to solve some of the problems encountered in CCD tasks, specifically those focused on the re-use of available information and the use of many different experts' experience. We presented a case study that follows our proposal in the Spanish normative for mid-level vocational education domain, implementing a competences and courses creation tool. As future work, we are focused in two different points: (i) to apply our methodology in several education domains and (ii) to enhance the Semantic Reasoning stage with Reflexive Ontologies [15] in order to enhance our domain ontology with quicker answers for the queries [16].

References

1. Swedish national Agency for Higher Education. e-Learning Quality: Aspects and Criteria for Evaluation of e-Learning in Higher Education. Report 2008:11 R (2008)
2. Diamond, R.M.: Designing and Assessing Courses and Curricula. Jossey-Bass, San Francisco (1998)
3. Aammodt, A., Plaza, E.: Case-Based Reasoning: Foundational issues, methodological variations, and system approaches. *AICom - Artificial Intelligence Communications* 7(1-3), 39–59 (1994)
4. Noy, N.F., McGuinness, D.L.: *Ontology Development 101: A Guide to Creating Your First Ontology*. Stanford Medical Informatics Technical Reports SMI-2001, 880 (2001)
5. Sanin, C., Szczerbicki, E.: Set of Experience: A Knowledge Structure for Formal Decision Events. *Foundations of Control and Management Sciences Journal* 3, 95–113 (2005)

6. Leake, D.: CBR in Context: the Present and Future. In: Case Based Reasoning: Experiences, Lessons and future Directions, pp. 3–30. AAI/MIT Press, Menlo Park (1996)
7. Bréal, M.: *Essai de sémantique*. Hachette, Paris (1897)
8. Gruber, T.R.: Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human-Computer Studies* 43(5-6), 907–928 (1995)
9. Rector, A., Drummond, N., Horridge, M., Rogers, J., Knublauch, H., Stevens, R., Wang, H., Wroe, C.: OWL Pizzas: Practical Experience of Teaching OWL-DL: Common Errors & Common Patterns. In: Motta, E., Shadbolt, N.R., Stutt, A., Gibbins, N. (eds.) EKAW 2004. LNCS (LNAI), vol. 3257, pp. 63–81. Springer, Heidelberg (2004)
10. Sirin, E., Parsia, B., Grau, B.C., Kalyanpur, A., Katz, Y.: A Practical OWL-DL Reasoner. *Journal of Web Semantics* 5(2), 51–53 (2007)
11. Toro, C., Sanín, C., Vaquero, J., Posada, J., Szczerbicki, E.: Knowledge Based Industrial Maintenance Using Portable Devices and Augmented Reality. In: Apolloni, B., Howlett, R.J., Jain, L. (eds.) KES 2007, Part I. LNCS (LNAI), vol. 4692, pp. 295–302. Springer, Heidelberg (2007)
12. Díaz-Agudo, B., González-Calero, P.A., Recio-García, J.A., Sánchez, A.: Building CBR systems with jCOLIBRI. *Science of Computer Programming* 69(1-3), 68–75 (2007)
13. Knublauch, H.: The Protégé-OWL API, <http://protege.stanford.edu/plugins/owl/api/index.html> (last visited on April 4, 2009)
14. Sanin, C., Szczerbicki, E., Toro, C.: An OWL Ontology of Set of Experience Knowledge Structure. *Journal of universal Computer Science* 13(2), 209–223 (2007)
15. Toro, C., Sanin, C., Szczerbicki, E., Posada, J.: Reflexive Ontologies: Enhancing ontologies with Self-Contained Queries. *Cybernetics and Systems: An International Journal* 39(2), 171–189 (2008)
16. Cobos, Y., Toro, C., Sarasua, C., Vaquero, J., Linaza, M.T., Posada, J.: An Architecture for Fast Semantic Retrieval in the Film Heritage Domain. In: *Proceedings of the 6th International Workshop on Content-based Multimedia Indexing*, London, UK, pp. 272–279 (2008)

Using the Mesh Thesaurus to Index a Medical Article: Combination of Content, Structure and Semantics

Jihen Majdoubi, Mohamed Tmar, and Faiez Gargouri

Miracl Laboratory, ISIMS Institute, BP 1030-3018, Sfax, Tunisia
majdoubi_jihene@yahoo.fr, mohamed.tmar@isimsf.rnu.tn,
Faiez.Gargouri@fsegs.rnu.tn

Abstract. This paper proposes an automatic method using a MeSH (Medical Subject Headings) thesaurus for generating a semantic annotation of medical articles. First, our approach uses NLP (Natural Language Processing) techniques to extract the indexing terms. Second, it extracts the Mesh concepts from this set of indexing terms. Then, these concepts are weighed based on their frequencies, locations in the article and their semantic relationships according to MeSH. Next, a refinement phase is triggered in order to upgrade the frequent ontology's concepts and determine the ones which will be integrated in the annotation. Finally, the structured result annotation is built.

Keywords: Semantic annotation, medical article, Mesh thesaurus.

1 Introduction

In the medical field, scientific articles represent a very important source of knowledge for researchers of this domain. These researchers usually need to deal with a large amount of scientific and technical articles for checking, validating and enriching their research work. But due to the large volume of scientific articles published on the web, an efficient detection and use of this knowledge is quite a difficult task. The semantic web can facilitate this task: it can be carried out by associating to each scientific article a semantic annotation. This annotation provides a more precise description of the knowledge contained in the document in order to facilitate the exploitation and the access of this document. The automatic indexing of the biomedical literature using MeSH indexing terms has been investigated by several researchers working on text in English as well as other European languages such as French, German and Portuguese. Indexing approaches are generally based on the content of the document or a combination of content and structure such as [10] which assigned an additional weight to the terms that are extracted from the title or the subtitle of the document. However, to our knowledge due to the large number of indexing terms and even larger number of possible combinations most of these efforts. Indeed, to determine a term importance, most semantic annotation approaches are based on the statistical measure. So, any consideration of the term semantic in the calculation of its weight is taken into account. For example, the term weight is calculated independently of its synonym occurrences. This lack motivated us to explore an annotation approach that calculates

term importance based on its frequencies, its locations in the article and its semantics relationships. In this paper we present a novel weighing technique by intuitively interpreting the conceptual information in the Mesh thesaurus. To determine term importance, this technique exploits its relationships in Mesh, rather than relying only on a statistical measure.

The remainder of this paper is organized as follows. Section 2 presents the Mesh thesaurus. Section 3 details our annotation approach. Finally, section 4 summarizes our proposal and outlines future work related to our annotation methodology as a whole.

2 Mesh Thesaurus

The language of biomedical texts, like all natural language, is complex and poses problems of synonymy and polysemy. Therefore, many terminological systems have been proposed and developed such as Galen [8], Gene Ontology [1] and Mesh [5]. In our context, we have chosen Mesh because it meets the aims of medical librarians and it is a widely used tool for indexing literature.

In our approach, Mesh is characterized by a set of concepts ($C=\{c_1, c_2, \dots, c_n\}$) and a set of relationships (R) between these concepts.

Each concept ($c_i=(name_{ci}, \{s_{i1}, s_{i2}, \dots, s_{in}\}) \in C$) consists of a set of synonym terms; one of them gives its name to the concept. Each term can be either a simple or a composed term and it is represented by a list of attributes where each attribute is a single term ($s_{ij}=(att_{ij1}, att_{ij2}, \dots, att_{ijk})$).

For example, $c_1=(infection\ of\ ear, \{otit, infection\ of\ ear, infection\ bacterial\ of\ ear, inflammation\ of\ ear, inflammation\ bacterial\ of\ ear\})$, is expressed by $((infection, ear), \{(otit, (infection, ear), (infection, bacterial, ear), (inflammation, ear), (inflammation, bacterial, ear)\})$.

Formally, a relationship (R) is a set of triplets (c_i, c_j, r) where c_i and c_j are two concepts related by $r \in \{hypernymy, hyponymy, meronymy, holonymy\}$, thus $R \subset C \times C \times \{hypernymy, hyponymy, meronymy, holonymy\}$

- *hypernymy*: the generic concept used for the is-a relation (cancer is *hypernym* of Cancer of blood), thus (cancer, cancer of blood, hypernymy).
- *hyponymy*: the specific concept used for the is-a relation (Cancer of blood is *hyponym* of cancer), thus (cancer of blood, cancer, hyponymy).
- *holonymy*: the global concept used for the *has-a* relation (face *has-a* {eyes, nose}, face is *Holonym of* eyes), thus (face, eyes, holonymy).
- *meronymy*: the concept which is part of another concept (nose is *meronym* of face), thus (eyes, face, meronymy).

3 Our Approach

Our objective is to develop an approach which, starting from a medical article and the Mesh thesaurus, generates a semantic annotation that describes the article content.

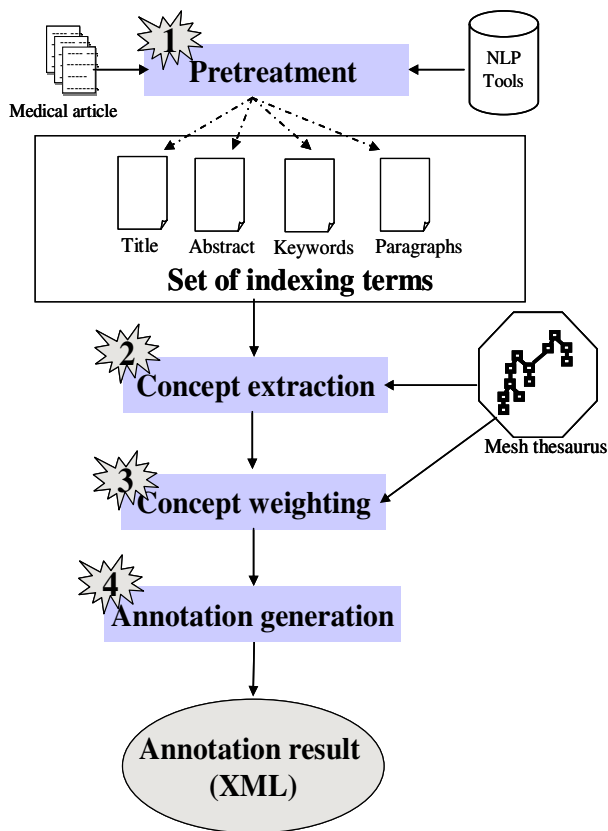


Fig. 1. General architecture of our annotation method

Our proposed annotation methodology as schematized in Figure 1, consists of four main steps. At the first step (Pretreatment), being given an article and the NLP tool GATE [2], the system extracts the set of indexing terms from this article. At step 2, this set is used in order to extract the Mesh concepts. In step 3, the extracted concepts are then weighed. Finally at step 4, an annotation result is built.

3.1 Pretreatment

As mentioned previously, the document structure will be used to weigh its terms. Indeed, a term that appears in the title will have a higher weight than a term that appears in the abstract. Thus, in this step the system extracts the various document tags (title, abstract, keywords and paragraphs). After, the lemmatisation of these tags is ensured by TreeTagger in order to generate a set of indexing terms existing in the document to be annotated. For further details about the *pretreatment*, the reader is referred to [6].

3.2 Concept Extraction

This step consists of finding the different Mesh concepts existing in the indexing terms generated by the pretreatment step. In [7], to extract the Mesh concepts we have used the vector space model [9] which measures the documents similarities with the query. We have adapted this model by substituting the document by the sentence and the query by the terms of a Mesh concept.

Each stemmed sentence S is a list of stems ordered in S as they appear in the text. Let $S=(t_1, t_2, \dots, t_n)$.

Each concept c_i is processing with TreeTagger in order to return its stemmed form for each one of its terms s_{ij} ($s_{ij}=(att_{ij1}, att_{ij2}, \dots, att_{ijk})$).

$$\text{Thus, } Sim_s(S, s) = \max(k) \tag{1}$$

$$\begin{aligned}
 & i \in \{1, 2, \dots, n\}, j \in \{1, 2, \dots, m\} \\
 & k \in \{1, 2, \dots, \text{Min}(n-i, m-j)\} \\
 & (t_i \dots t_{i+1} \dots t_{i+k}) = (att_j \dots att_{j+1} \dots att_{j+k})
 \end{aligned}$$

For a concept c composed of n terms, its similarity $sim_c(S, c)$ in a sentence S is defined as follows: $Sim_c(S, c_i=(name_{c_i}, \{s_{i1}, \dots, s_{in}\})) = \sum_{s \in \{s_{i1}, s_{i2}, \dots, s_{in}\}} Sim_s(S, s)$ (2)

Example

Let a sentence $S=$ “*otit is an inflammation that attacks ear*” and a concept $c_i=(infection\ of\ ear, \{otit, infection\ of\ ear, infection\ bacterial\ of\ ear, inflammation\ of\ ear, inflammation\ bacterial\ of\ ear\})$, the calculation of $sim_c(S, c_i)$ is performed as follows.

After the step of pretreatment, $S=$ ”otit inflammation ear” and c_i is expressed by ((infection, ear), {(otit), (infection, ear), (infection, bacterial, ear), (inflammation, ear), (inflammation, bacterial, ear)}).

$Sim_s(S, otit)=1, Sim_s(S, infection\ ear)=0, Sim_s(S, infection\ bacterial\ ear)=0, Sim_s(S, inflammation\ ear)=2$ and $Sim_s(S, inflammation\ bacteria\ ear)=1$. Thus $Sim_c(S, c_i)=4$.

The concept extraction process is iterative. For each stemmed sentence, we compute its similarity with the set of Mesh concepts. Only those having a positive score are proposed to the system like Candidate Concept of the sentence S ($CC(S)$).

$$name_{c_i} \in CC(S) \text{ if } \exists c_i \in C, Sim_c(S, c_i) > 0 \tag{3}$$

However, a sentence can contain several $CC(S)$, the problem here is which will be the Best Candidate Concept ($BCC(S)$) among the set of $CC(S)$. Two cases are dealt with in this situation.

- Case1: disjoint concept: this case occurs when the concept c has no any common attributes with any concept of the set of $CC(S)$, thus c will be systematically added to $BCC(S)$. Let us consider for each concept, its set of attributes. For example, for the concept $c_i=(infection\ of\ ear, \{otit, infection\ of\ ear, infection\ bacterial\ of\ ear, inflammation\ of\ ear, inflammation\ bacterial\ of\ ear\})$, the set of attributes of concept c_i is done as follows: $Attributes(c_i)=\{otit, infection, ear, bacterial, inflammation\}$.

$$\text{if } \exists c \in CC(S), \forall c' \in CC(S) - \{c\}, Attributes(c) \cap Attributes(c') = \emptyset \text{ then } c \in BCC(S) \tag{4}$$

- Case2: joint concept: this case occurs when a concept c has one or more common attributes with at least the one other concept of the set of $CC(S)$. Thus, we return the concept which has the highest similarity for each sentence.

$$\begin{aligned} & \text{if } \exists c \in CC(S), c' \in CC(S), Attributes(c) \cap Attributes(c') \neq \emptyset \\ & \text{then } \arg \max_{\substack{\forall c' \in CC(S), \\ Attributes(c) \cap Attributes(c') \neq \emptyset}} \|Sim_C(c', S)\| \in BCC(S) \end{aligned} \quad (5)$$

Finally, we generate the $BCC(D)$ as follows:

$$BCC(D) = \bigcup_{S_i \in D} BCC(S_i) \quad (6)$$

These extracted concepts ($BCC(D)$) are then weighed in order to determine the importance of each one in the document and select the concepts that will belong to the annotation result.

3.3 Concept Weighing

Given a set of extracted concepts ($BCC(D)$), issued from the step of *concept extraction*, we calculate the concept's weight by using two measures: the Content Structure Weight (CSW) and the Semantic Weight (SW).

3.3.1 Content Structure Weight

We can notice that the frequency is not a main criterion to calculate the CSW of the concept c . Indeed, the CSW takes into account the concept frequency and especially the location of each one of its occurrences. For example, the concept of the "Title" receives a high importance (*10) compared to the concept of the "Paragraphs" (*2). Table 1 shows the various coefficients used to weigh the concept locations. These coefficients were determined in an experimental way in [4] and also used in [3].

$$CSW(c, D) = \sum_{c \in A} f(c, D, A) \times W_A \quad (7)$$

Where: $f(c_i, D, A)$: the occurrence frequency of the concept c_i in document D at location A with $A \in \{\text{title, keywords, abstract, paragraphs}\}$,

W_A : weight of the position A (see Table 1),

Table 1. Weighing coefficients

Concepts location	Weight of the location
Title	10
Keywords	9
Abstract	8
Paragraphs	2

3.3.2 Semantic Weight

The Semantic Weight of the concept c depends on its *hypernyms*, *hyponyms*, *meronyms* and *holonyms* existing in the $BCC(D)$.

To do so, we use the function $Relatedto_E$ that associates for a given concept c , and a relationship r , all concepts satisfying r among the set of concepts E .

$$relatedto_E : \begin{matrix} C \times R & \rightarrow & 2^E \\ (c, r) & \mapsto & \{c' \in E, (c, c', r) \in R\} \end{matrix}$$

To calculate the $SW(c_i, D)$, we apply the following formula:

$$SW(c, D) = \frac{\sum_{g \in relatedto_{BCC(D)}(c, r)} CSW(g, D) \times W_r}{|relatedto_{BCC(D)}(c, r)|} \tag{8}$$

With W_r is the weight that measures the importance of each one of these relationships (*hypernymy*, *hyponymy*, *meronymy* and *holonymy*).

For a given concept c_i , we have on the one hand its Content Structure Weight ($CSW(c_i, D)$) and on the other hand its Semantic Weight ($SW(c_i, D)$), its concept frequency ($cf(C, D)$) is determined as the sum of the both.

$$\text{We compute } cf(C, D) \text{ as } CSW(C, D) + SW(C, D). \tag{9}$$

Once the concept Frequency is calculated, it is used to calculate the weight of a concept c in a D document ($W(c, D)$).

$$W(c, D) = cf(c, D) \cdot \ln\left(\frac{N}{df}\right) \tag{10}$$

Where N is the total number of documents and df (document frequency) is the number of documents which concept c_i occurs in.

Once the (FC(D)) frequent concepts (concepts with a weight that exceeds the threshold) are extracted from a document D , they are used to build the annotation result.

3.4 Annotation Generation

Each concept of $FC(D)$ can have several means because the Semantic Weight of a concept is majored by the frequency of its *hypernyms*, *hyponyms*, *meronyms* and *holonyms* concept. Consequently, the size of the annotation result increases and its exploitation becomes increasingly tedious. Thus, a refinement phase is triggered in order to upgrade the set of $FC(D)$ and select the Valid Concepts ($VC(D)$). For this refinement we have defined the set of rules.

- *Strong Concept* is a concept that belongs to $FC(D)$ and it has at least the half of its hyponyms(or meronyms) in the set of $FC(D)$.

$$\exists c \in FC(D), \frac{|relatedto_{FC(D)}(c,r)|}{|relatedto_C(c,r)|} \geq 0.5 \Rightarrow StrongConcept(c) \quad (11)$$

with $r \in \{hyponymy, meronymy\}$.

For each c of $FC(D)$, if c is a strong-concept, then c belongs to $VC(D)$. Indeed, a document that contains the concepts football, volleyball, basketball, handball, is a document that generally represents sport.

$$\exists c \in FC(D), StrongConcept(c) \Rightarrow c \in VC(D) \quad (12)$$

– *WeakConcept* is a concept that belongs to $FC(D)$ and it has less of the half of its hyponymy(or meronymy) in the set of $FC(D)$.

$$\exists c \in FC(D), \frac{|relatedto_{FC(D)}(c,r)|}{|relatedto_C(c,r)|} < 0.5 \Rightarrow WeakConcept(c) \quad (13)$$

with $r \in \{hyponymy, meronymy\}$.

For each c of $FC(D)$, if c is a *WeakConcept*, then its hyponyms(or meronyms) belongs to $VC(D)$.

$$\exists c \in FC(D), WeakConcept(c) \Rightarrow \begin{matrix} c' \in VC(D) \\ (c,c',hyponymy) \in R \\ (c,c',meronymy) \in R \end{matrix} \quad (14)$$

– *SingleConcept* is a concept that belongs to $FC(D)$ and it has neither hyponym nor meronym and it is neither hypernym nor holonym.

$$\exists c \in FC(D), \frac{|relatedto_{FC(D)}(c,r)|}{|relatedto_C(c,r)|} = 0 \Rightarrow SingleConcept(c) \quad (15)$$

With $r \in \{hypernymy, hyponymy, meronymy, holonymy\}$.

Each *single-concept* is assigned to the $VC(D)$.

$$\exists c \in FC(D), SingleConcept(c) \Rightarrow c \in VC(D) \quad (16)$$

4 Conclusion

The work developed in this paper outlined an automatic approach for representing the semantic content of medical articles using a Mesh thesaurus.

In this paper, we have proposed a new concept weighing technique that intuitively interprets Mesh's conceptual information to calculate term importance.

This approach is independent of any domain and can be generalized to all ontology or thesaurus. While the system of annotation is implemented, we are currently working on the evaluation of our method on a medical corpus in order to measure its relevance.

References

1. Ashburner, M., Ball, C., Blake, J., Butler, H., Cherry, J., Corradi, J., Dolinski, K., Janan, T., Eppig, T., Harris, M.: Creating the Gene Ontology resource: design and implementation. In: *Genome Research*, pp. 1425–1433 (2001)
2. Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V.: GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In: *ACL 2002* (2002)
3. Desmontils, E., Jaquin, C.: Indexing a Web site with a terminology oriented ontology. In: *The Emerging Semantic Web*, pp. 181–197. IOS Press, Amsterdam (2002)
4. Gamet, J.: Indexation de pages web, Report of DEA université de Nantes (1998)
5. Humphreys, B., Lindberg, D.: The UMLS project: making the conceptual connection between users and the information they need. *Bulletin of the Medical Library Association* 81(2), 170 (1993)
6. Majdoubi, J., Tmar, M., Gargouri, F.: Indexing a medical article with a Terminology Oriented Ontology. In: *ICTA 2009: International conference Information and Communication Technologies and Accessibility*, Hammamet, Tunisie (May 2009)
7. Majdoubi, J., Tmar, M., Gargouri, F.: Ontology-Based semantic annotations of medical articles. In: *SEKE 2009: the 21st International Conference on Software Engineering and Knowledge Engineering*, Boston, USA, July 1-3 (2009)
8. Rector, A., Rogers, J., Pole, P.: The GALEN High Level Ontology. In: *Fourteenth International Congress of the European Federation for Medical Informatics, MIE 1996*, Copenhagen, Denmark (1996)
9. Salton, G.: *The SMART Retrieval System: Experiment in Automatic Document Processing*. Prentice-Hall, Englewood Cliffs (1970)
10. Soualmia, L., Golbreich, C., Darmoni, S.: Representing the MeSH in OWL: Towards a semi-automatic Migration. In: *First International Workshop on Formal Biomedical Knowledge Representation, collocated with KR 2004*, Whistler, Canada, pp. 1–12 (2004)

Building Decision Trees to Identify the Intent of a User Query

Marcelo Mendoza^{1,3} and Juan Zamora²

¹ Yahoo! Research, Santiago, Chile

² Applied Computational Intelligence Lab (INCA), Department of Informatics,
Universidad Técnica Federico Santa María, Chile

³ Computer Science Department, Universidad de Valparaíso, Chile

Abstract. In this work we explore the use of decision trees to identify the intent of a user query, based on informational, navigational, and transactional categorization. They are based on decision trees, using the C4.5 implementation. The classifier will be built from a query data set larger than any previously used, allowing the conclusions to have a greater reach. Unlike previous works, we will explore features that have not been evaluated before (e.g. PageRank) combined with features based on text and/or click-through data. The results obtained are very precise and the decision tree obtained allows us to illustrate relations among the variables used for classification determining which of these variables are more useful in the classification process.

1 Introduction

At present, the Web is the largest and most diverse document database in the world. To access the content, Web users use search engines, where they formulate their queries. Then, the most relevant sites and pages are displayed for the user as a list of answers ordered by relevance to the query. Given the vastness of the Web and the little information on which a search engine has to base the query, the answers lists tend to be imprecise.

Broder [2] suggested that a way to improve the precision of the answers lists was to distinguish between the query and the needs of the user that formulates the query (the intent of a user query). That way it would be possible to use ranking algorithms that are adjusted to the type of user need.

Broder identified three types of needs from the queries. The first type is *Informational*, in which users search for information available in the content of the sites / pages. The most common form of interaction with this type of content is reading. Second is the *Navigational* need, where users search for a specific site whose URL they do not remember. The third type of need is *Transactional*, in which the users search for a page / site to make some kind of transaction such as download a file, make plan reservations, buy / sell, etc.

Later, numerous works concentrated on identifying features of the queries that allow them to be categorized into the Broder taxonomy. These classifiers seek to automatically classify the queries in the categories described earlier. There has

been a wide range of results, and in general, there are no conclusive results. This is largely because these classifiers have been evaluated using very small data sets or under experimental conditions that make it impossible to generalize the obtained results.

2 Related Work

Once Broder proposed the web search taxonomy, defining the informational, navigational, and transactional categories, Rose and Levinson [10] extended Broder's categories. They identified types of frequent interactions between users and the recommended sites / pages, depending on the type of query posed.

Following this line of investigation, Kang and Kim [4] proposed characterizing the queries based on the distribution of the terms they contain. Based on a set of queries pulled from the TREC [1] collection and classified by experts into the Broder categories, they obtained a collection of terms frequently used to pose navigational or informational queries. Using mutual information criteria between the query terms and the titles and snippets from the selected pages / sites in the query sessions, they were able to automatically classify the queries with nearly 80% precision, on an evaluation data set of 200 queries.

Later, Lee *et al.* [5] proposed identifying the type of query by observing the levels of bias in click distributions in the query sessions being classified. Intuitively, an informational query should have a distribution with more clicks than a navigational query, where it is presumed that user preferences will generally be concentrated on just one site. From this idea, they proposed a classifier that reached near 50% precision on an evaluation data set of 50 queries.

A similar strategy was proposed by Liu *et al.* [6] who introduced two measurements based on click-through data to characterize queries based on their session logs in the search engine: **nRS** (number of sessions in which clicks are registered before a given position n) and **nCS** (number of sessions registering less than n clicks). Using a decision tree they were able to reach close to 80% precision on 400 manually classified queries. Later, Baeza-Yates *et al.* [1] proposed using a query vector representation based on text and click-through data. Using techniques such as Support Vector Machines (SVMs) and Probabilistic Latent Semantic Analysis (PLSA), they reached almost 50% precision on a dataset of 6,000 queries semi-automatically classified into the Broder categories (the vector representations were clustered and then those clusters were labeled).

Recently, using classifiers based on features extracted from the query session registry, such as the number of query terms, number and distribution of clicks, among others, Jansen *et al.* [3] achieved nearly 74% precision on an evaluation data set of 400 queries manually classified by a group of experts. Finally, on the same data set used in the Baeza-Yates *et al.* [1] experiments, Mendoza and Baeza-Yates [7] showed that the features based on text or on combinations of

¹ Text Retrieval Conference co-sponsored by NIST. Dataset available on <http://trec.nist.gov/data.html>

text and click-through data differentiated more queries in the Broder taxonomy than those based just on click-through data.

3 The Classifier

3.1 Data Set

Based on a query log file provided by AOL², consisting of a three-month period in 2006, which contains 594,564 queries associated to 765,292 query sessions, registering 1,124,664 clicks over 374,349 selected pages / sites, we will analyze the features that will be most useful for the classification process. Experts from the *Applied Computational Intelligence Lab (INCA)* of Chile and the *Universitat Pompeu Fabra* of Spain have collaborated in the manual classification of 2,000 queries randomly pulled from the AOL log. The experts completed a survey similar to that used by Broder in his first manual categorization experiment [2]. The final set of queries used for the analysis was compiled from those queries where the answers of the experts were in agreement. As a result, 1,953 queries were labeled by consensus, discarding only 2% of the initial data set. The queries categorized by consensus were distributed from greater to lesser proportion in the informational, navigational, and transactional classes, with 52%, 33%, and 15% over the complete data set, respectively.

3.2 Feature Analysis

Based on the features analyses by Lee *et al.* [5], by Liu *et al.* [6] and by Mendoza and Baeza-Yates [7], we will use the following features for the classification process, selected according to their discriminatory capacity:

- Number of terms in the query (**nterms**): The number of words that compose each query. Mendoza and Baeza-Yates [7] show that a significant proportion of the queries with five or more query terms belongs to the informational category.
- Number of clicks in query sessions (**nclicks**): Average number of selected documents per session calculated over the set of query sessions related to a query (the sessions where the query was formulated). Lee *et al.* [5] show that a significant number of navigational queries concentrates only a few clicks per session.
- Levenshtein distance: distance function calculated among the terms that compose the query and the snippets (the snippet is compounded by the excerpt presented with the query result, the title and the URL of the selected document). Mendoza and Baeza-Yates [7] show that the distance distribution for the informational category has a media of 39.67 calculated over 12,712 pairs queries - snippets. The distribution gotten for non-informational queries (navigational and transactional queries) has a media of 37.6 over 10,540 pairs queries - snippets.

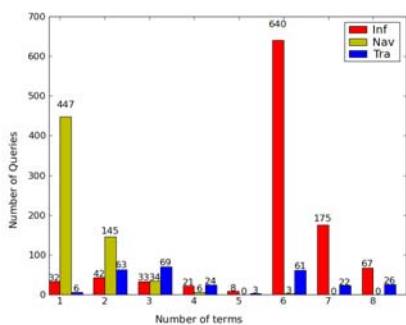
² America On-Line Search Engine. Query log files are available for research purposes on <http://www.gregsadetsky.com/aol-data/>

- Number of sessions with less than n clicks over the total of sessions associated to a query (**nCS**): Liu *et al.* [6] introduce the **nCS** feature, that is defined as the number of sessions of a query q that register less than n selections, calculated over the set of sessions where q was formulated. Liu *et al.* calculate the measure for $n = 2$ and $n = 3$. In the extremes of the distribution gotten, this is, for values around 0.95 and 0.05, the informational category exceed the values achieved by non-informational queries.
- Number of clicks before the n -th position of the query ranking (**nRS**): In [6] Liu *et al.* introduce the **nRS** feature, that is defined as the number of sessions of a query q that register selections only in the top- n results of the answer list of q , calculated over the set of sessions where q was formulated. Liu *et al.* calculate the measure for $n = 5$ and $n = 10$. In the extreme of both distributions, in the class of mark 0.95, non-informational queries exceed informational queries.
- PageRank [8]: Mendoza and Baeza-Yates [7] analyze the hyperlink structure between the pages / sites selected for each category. To do this, they calculate the PageRank³ measure considering the collection of selected documents in each query category as a subgraph of linked pages / sites. Then, they calculate over each category collection the PageRank measure using a fixed-point algorithm over the matrix of hyperlinks. The authors show that the PageRank values for documents selected in sessions of non-informational queries are higher than the ones selected in informational queries.

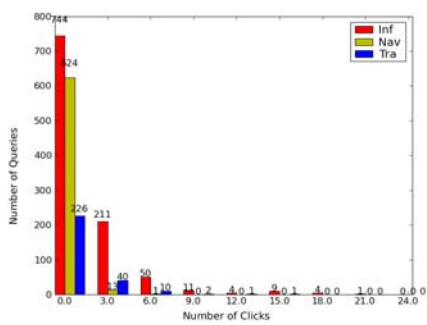
In order to illustrate the discriminatory capacity of each of the features considered, we will plot the distribution of the characteristic according to the Broder categorization using our dataset. In the case of the **nCS** and **nRS** features, we have tested versions 2CS, 3CS, 5RS and 10RS, which showed the best results in Liu *et al.* [6]. According to the distributions obtained, the variables that have a greater discriminatory capacity are 2CS and 5RS. They are shown along with the results for **nterms**, **nclicks**, Levenshtein distance and Page Rank, in Figure 1.

As we can observe in Figure 1a) the navigational queries generally have fewer terms than the informational queries. The behavior of this characteristic is not as clear for the transactional class. Figure 1b) shows that some informational queries register more than 9 different sites / pages selected in their sessions. This usually does not occur in the case of navigational or transactional queries. Figure 1c) shows that in general, the Levenshtein distance calculated between query terms and snippets is less in the case of navigational queries than for the other categories. Figure 1d) illustrates that a good amount of informational queries register clicks in pages / sites with low Page Rank, as opposed to transactional or navigational queries. Figure 1e) shows that the characteristic 2CS is useful for differentiating

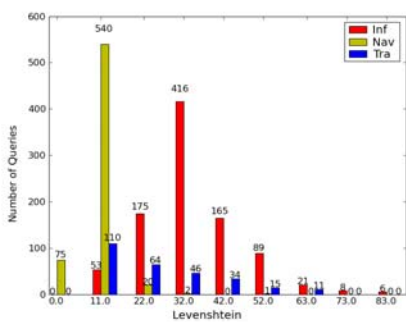
³ The PageRank of a page / site is the stationary probability of visiting it in a random walk of the web where the set of states of the random walk process is the set of pages / sites of the web and the transitions between states are one of the following two cases: a) To follow an outgoing link of the page, b) To jump to another page / site selected randomly from the entire web.



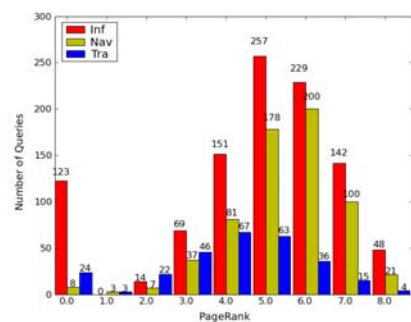
(a)



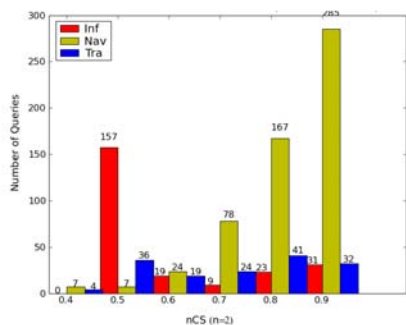
(b)



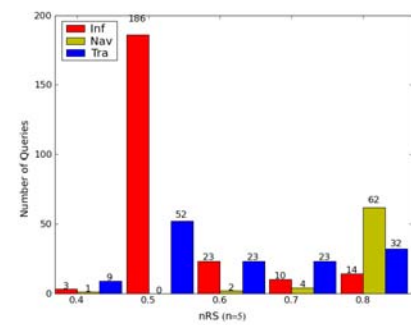
(c)



(d)



(e)



(f)

Fig. 1. Feature analysis by category: (a) nterms, (b) nclicks, (c) Levenshtein distance, (d) PageRank, (e) 2CS and (f) 5RS

between navigational queries. Figure 1f) shows something similar for the 5RS characteristic, which in this case is useful for differentiating informational queries.

3.3 Building the Decision Tree

To build the decision tree we use 30% of the data set for evaluation and 70% for training (randomly divided). We use C4.5 to built the decision tree. C4.5 [9], widely used for statistical classification, chooses the feature with the highest normalized information gain to split the training data. Then the algorithm recurses on each training data subset until no more features can be used. The C4.5 implementation corresponds to that provided by the open-source Weka software⁴. The tree was built by Weka in just 0.12 seconds and is shown in Figure 2.

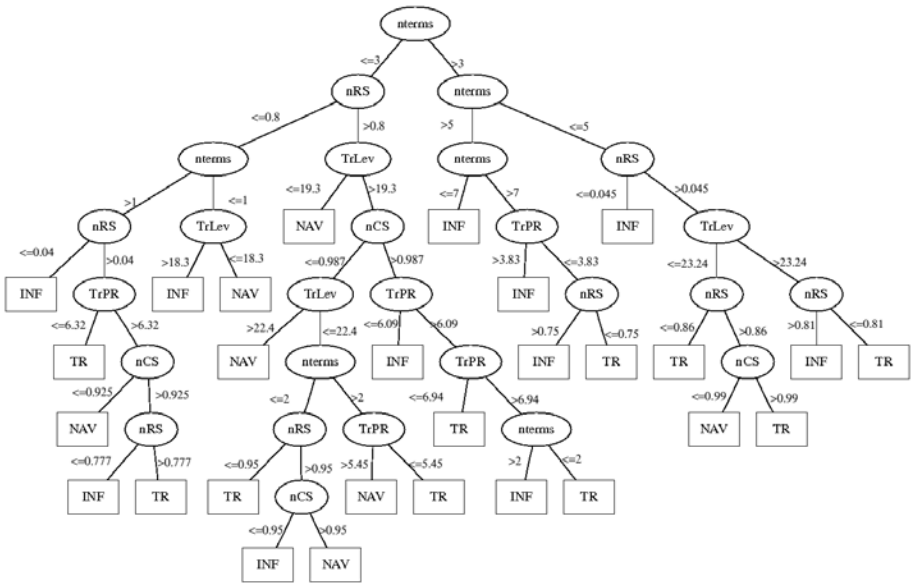


Fig. 2. Decision tree proposed for the problem of query intent identification

We can see that a characteristic relevant to the majority of tree branches is **nterms**, which creates the first partition in the data set. This characteristic can detect informational queries when $3 < \text{nterms}$. Combinations with **5RS** can detect the transactional class in this branch. A third case in this branch identifies navigational queries, for which the **2CS** characteristic also needs to be used. When $\text{nterms} \leq 3$, more features are needed to be able to classify. For example, PageRank and Levenshtein distance are used.

⁴ Waikato Environment for Knowledge Analysis (Weka), University of Waikato, New Zealand. Available on <http://www.cs.waikato.ac.nz/ml/weka>

4 Evaluation

To consider the costs of evaluation (tradeoff predictive / discriminative) from the comparison of the nominal / predicted class, we consider the four possible cases: true positives (tp), false positives (fp), false negatives (fn) and true negatives (tn). Based on these four cases, the following performance evaluation measurements will be calculated: Precision ($\frac{tp}{tp+fp}$), FP rate ($\frac{fp}{fp+tn}$), TP rate or Recall ($\frac{tp}{tp+fn}$) and F-measure ($\frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}}$). In the case of the F-measure, the F_1 version has been considered, which corresponds to the harmonic mean between Precision and Recall, given that it can assess the commitment between both criteria.

The classifier performance will be evaluated by comparing the results from each category analyzed (comparing reference class vs. other classes). The results of this analysis are shown in Table 1.

Table 1. Performance evaluation of the proposed decision tree

Comparison	Measures			
	FP Rate	Precision	Recall	F-Measure
(1) Informational - Other Classes	0.182	0.841	0.917	0.878
(2) Navigational - Other Classes	0.066	0.876	0.953	0.913
(3) Transactional - Other Classes	0.032	0.673	0.355	0.465
(4) Weighted Average	0.120	0.826	0.840	0.824

As we can see in Table 1, the results are high in both precision and in recall, with an average of 0.824 for the F_1 -measure. The lowest performing class is transactional, even though this class obtains a high precision over the portion of queries in this category.

To evaluate the predictive / discriminative tradeoff between unbalanced classes, we use ROC curves (**R**eceiver **O**perating **C**haracteristics), that means to plot TP Rate (benefit) vs. FP Rate (cost). To the extent that the classifier performs well, the area under the curve (AUC) will be greater (maximizing the cost / benefit relationship). The AUC values obtained were as follows: 0.8979, 0.9551 and 0.7357, for the informational, navigational, and transactional categories, respectively. The results obtained show that the predictive / discriminative tradeoff is nearly optimal in the case of navigational and informational categories, and it is acceptable in the transactional case.

5 Conclusions

In this work we have presented a new query classifier according to the Broder taxonomy, based on features and built using decision trees through C4.5 implementation. The resulting tree allows precedence relationships between features to be established. Thus, it can be concluded that all the features considered

have been relevant for identifying the Broder categories. Experimental results affirm that the resulting classifier obtains both high precision and recall results, maintaining a balance in the prediction / discrimination relationship, especially for the informational and navigational categories. Given the low performance obtained regarding the transactional category, it has been more challenging to identify features that lead to a clear detection of this type of queries.

For future work, the exploration of new features and / or classification techniques should be considered so as to improve the results for the transactional category.

Acknowledgments

Dr. Mendoza was partially supported by DIPUV project 52/07 from Universidad de Valparaiso, Chile. Mr. Zamora was supported by a fellowship for scientific initiation of the Graduate School of the UTFSM, Chile. Finally, we would like to thank the contributions of Ricardo Baeza-Yates, Hector Allende-Cid, Libertad Tansini and Katharine Sherwin.

References

1. Baeza-Yates, R., Calderón-Benavides, L., González-Caro, C.: The intention behind web queries. In: Crestani, F., Ferragina, P., Sanderson, M. (eds.) SPIRE 2006. LNCS, vol. 4209, pp. 98–109. Springer, Heidelberg (2006)
2. Broder, A.: A taxonomy of web search. SIGIR Forum 36(2), 3–10 (2002)
3. Jansen, B., Booth, D., Spink, A.: Determining the informational, navigational and transactional intent of Web queries. *Information Processing and Management* 44(3), 1251–1266 (2008)
4. Kang, I.-H., Kim, G.: Query type classification for web document retrieval. In: Proceedings of SIGIR 2003, Toronto, Canada, July 28th - August 1st, pp. 64–71. ACM Press, New York (2003)
5. Lee, U., Liu, Z., Cho, J.: Automatic identification of user goals in web search. In: Proceedings of WWW 2005, Chiba, Japan, May 10-14, pp. 391–400. ACM, New York (2005)
6. Liu, Y., Zhang, M., Ru, L., Ma, S.: Automatic query type identification based on click through information. In: Ng, H.T., Leong, M.-K., Kan, M.-Y., Ji, D. (eds.) AIRS 2006. LNCS, vol. 4182, pp. 593–600. Springer, Heidelberg (2006)
7. Mendoza, M., Baeza-Yates, R.: A web search analysis considering the intention behind queries. In: Proceedings of LA-WEB 2008, Vila Velha, ES, Brazil, October 28-30, pp. 66–74. IEEE Computer Society Press, Los Alamitos (2008)
8. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web. In: Proceedings of WWW 1998, Brisbane, Australia, pp. 161–172. ACM, New York (1998)
9. Quinlan, R.: C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo (1993)
10. Rose, D.E., Levinson, D.: Understanding user goals in web search. In: Proceedings of WWW 2004, May 17-20, pp. 13–19. ACM, New York (2004)

Ontology-Based Concept Indexing of Images

Rossitza Setchi¹, Qiao Tang², and Carole Bouchard³

¹ School of Engineering, Cardiff University, Cardiff CF24 3AA, UK

² Beijing Digital China Limited, Beijing100080, China

³ Laboratory of New Product Design and Innovation, Ecole Nationale

Supérieure des Arts et Métiers, 75013 Paris, France

Setchi@cf.ac.uk, Qiao.Tang@gmail.com,

Carole.Bouchard@paris.ensam.fr

Abstract. The search for inspirational images is an important part of creative design. When identifying inspirational materials, designers search semantic domains that are different from the target domain and use semantic adjectives in combination with the traditionally used keywords. This paper describes research conducted within the TRENDS project, which aimed at developing a software tool for the needs of concept cars designers. The goal was to assist them with the process of collecting inspirational images from various sectors of influence. The paper describes the ontology tagging algorithm developed to index the images in the TRENDS database using concepts from two ontologies: a generic ontology called *OntoRo*, and a domain-specific ontology *CTA* developed for the needs of the project. The paper presents the evaluation of the developed algorithm and suggests areas for further research.

Keywords: Concept indexing, semantic indexing, ontology-based annotation, semantic search, ontology tagging.

1 Introduction

The research described in this paper was conducted in the context of a large-scale collaborative research project called TRENDS, which involved partners from four European countries specialized in automotive design, content-based retrieval of images, search engines, semantic-based systems, human-computer interaction and software design. The aim of the project was to develop a software tool for the needs of the designers of concept cars, which would assist them in collecting inspirational images from a number of sectors of influence, and stimulate their creativity.

The interviews conducted with the designers at the start of the project [1] revealed that they define their searches using: (i) design-specific elements such as shape, texture and color; (ii) sectors of influence (e.g. ‘furniture’ or ‘automotive’); (iii) keywords such as ‘boat’ or ‘money’; and (iv) semantic adjectives such as ‘fresh’, ‘aggressive’ or ‘soft’.

The TRENDS software integrates all these features, which have been developed collaboratively by the partners. This paper focuses on one element of the developed solution: the semantic search using domain-specific adjectives employed in design.

The paper describes the algorithm developed to index the images in the TRENDS database using concepts from two ontologies: a generic ontology called *OntoRo* [2], and a domain-specific ontology for designers called *CTA* [3].

The paper is organized as follows. Section 2 presents related work in the area of semantic indexing and ontology-based annotation. Section 3 describes the ontology tagging algorithm developed, and its evaluation. Section 4 presents conclusions.

2 Background and Related Work

Much of the latest research in image retrieval is focused on the semantic gap between the low level image features used in content-based retrieval and the high level concepts used in queries [4]. Most approaches are based on keywords that either correspond to identifiable items describing the visual content of an image or relate to the context and the interpretation of that image. Advances in image analysis, object detection and classification techniques may facilitate the automatic extraction of the first type of keywords. However, as stated in [5], keywords belonging to the second category are unlikely to be automatically obtained from images. A particular challenging aspect in this context is dealing with concepts, which have no visual appearance in the images. Examples include concepts related to categories such as time, space, events and their significance, as well as abstract terms and emotions [6]. Such concepts could be extracted from annotations or from the text accompanying the image. Most advanced image retrieval approaches, including the one advocated by the TRENDS partnership, employ hybrid techniques, which combine the use of visual features and text annotations. Since the focus of this paper is on extracting concepts that may represent the *meaning* of an image, the rest of this section will discuss only approaches aimed at *extracting concepts* from text and using them to *index* images.

The phrase *concept indexing* is associated with several very different approaches. As a semantic indexing method, however, concept indexing is defined in [7] as “the analytic process of identifying instances (*entities*) and abstract ideas (*concepts*) within a text document, and linking them to ontological concepts”. Concept indexing therefore can be used both for representing a document using abstract terms, and for assigning concepts to specific words in documents. In the above definition, concept index is a machine understandable index of entities and concepts contained in a document collection. An entity is an identifiable and discrete instance existing in a text document, while a concept is an abstract or general idea inferred or derived from specific instances.

The concept indexing approach normally involves two steps: (i) extracting entities from unstructured text-based content using a language knowledge base, and (ii) identifying concepts with the help of a concept knowledge base. Once entities and concepts are isolated, they are used to build a concept index [2]. The semantic concepts could be extracted and identified by disambiguating words’ senses using linguistic repositories [8], generic ontologies [2], semantic repositories [9] or domain-dependent ontologies [10]. As highlighted in [9], linguistic repositories such as WordNet [11] do not capture the semantic relationships between concepts. On the other hand, semantic repositories such as Cyc [12] developed to capture and represent common sense, do not represent linguistic relationships (e.g. whether two concepts are synonyms), and

domain dependent repositories like the Gene Ontology [10] only represent certain aspects of the domain, not the complete domain.

Semantic annotation shares similarities with semantic document indexing as it aims to provide some formalization of the content of the documents as a prerequisite for more comprehensive management [13]. Semantic annotations can be added both to documents, and portions of documents. The annotations normally used utilize a controlled vocabulary and are linked to some semantic descriptions (dictionaries, lexicons, glossaries, or ontologies). For example, the SemTag algorithm [14] has been applied to a set of 264 million pages generating 434 million automatically disambiguated semantic tags. The lexicon used by SemTag contains 200,000 words. The accuracy of the information retrieval is 82.01%.

It must be noted, however, that different ontologies may not have the same degree of formality. Controlled vocabularies, thesauri, and taxonomies are some of the most lightweight ontology types that have been widely used in annotation. These forms of vocabularies are not strictly formal and the annotations produced using them are normally pointers to terms in the vocabulary, which can be used to improve search [15]. These vocabularies can be used to find synonyms, antonyms, hyponyms and hypernyms for any word included in them. In addition to informal dictionaries, heavyweight ontologies (axiomatised and formal ontologies) are employed to incorporate formal semantics in the description of documents' content [16]. However, most of the formal ontologies do not include the vast number of terms that a thesaurus has. As noted in [15], thesauri, controlled vocabularies, and heavyweight ontologies are complementary since the first two can be used to provide agreed terms in specific domains while the latter provides formal semantics and constraint evaluation.

Although it is difficult to compare the existing approaches in terms of their accuracy, it is worth noting the good results achieved by the two concept indexing approaches capable of processing large scale document collections at word level. The first study [8] uses three ontologies (WordNet, OpenCyc and SUMO) containing 4,115 concepts to index texts word by word. The ontology mapping accuracy is 96.2% while the accuracy of the ontology tagging is estimated to be between 60% and 70%. Similar to this approach, the concept indexing algorithm described in [7] supports part of speech tagging, word sense disambiguation and indexing using concepts from a rich general purpose ontology. The average accuracy reported ranges from 76.40% to 78.91%, with the highest accuracy achieved when 90% of the corpus was used for training [7].

3 Ontology Tagging

3.1 Ontologies

The ontology tagger *OntoTag* described in this section uses two ontologies: a generic ontology *OntoRo* [2], and a domain-specific ontology for designers *CTA* [3].

OntoRo is a general purpose ontology based on the Roget's Thesaurus [17]. A thesaurus is a collection of terms organized within an agreed structure and linked through semantic relationships. The Roget's Thesaurus is a well known resource mainly used to facilitate the expression of ideas and assist in literacy composition. The decision to

build and use *OntoRo* instead of employing WordNet was primarily based on the fact that WordNet is a linguistic rather than a semantic repository. In comparison, Roget's has a well established structure where the words/phrases are grouped by their meaning. In addition, *OntoRo* contains a larger number of words/phrases compared to WordNet. Currently, *OntoRo* includes 68,920 unique words and 228,130 entries, which are classified into 990 concepts, 610 head groups, 95 subsections, 39 sections and 6 top level classes.

The *CTA* ontology is a purpose-built ontology, which has been developed using the so called Conjoint Trend Analysis (CTA) method [3] defined by studying the earliest phases of cognitive design. Fundamental to the CTA method is the establishment of a value-function-attributes chain, which uses semantic adjectives to link the marketing and design worlds. It has been found that the same semantic adjectives are used by designers when working with images and sketching new design concepts. An example of such a value-function-attributes chain is the sequence "comfortable life" (value) – "ergonomic" (semantic adjective) – "soft" (functional attribute). The *CTA* ontology is developed in Protégé by creating instances and linking them using abstraction, aggregation and dependency-based semantically-rich relations. Currently, the *CTA* ontology contains 10 classes and 503 instances.

3.2 Concept Indexing Algorithm

The concept indexing algorithm employed by the ontology tagger *OntoTag* is based on the assumption that there is a semantic link between an image and the text around it. The algorithm involves three steps.

Step (i): Retrieving web pages by targeted crawling and creating a collection of documents and images. This involves grabbing pages from web sites and domains identified by the designers as their sources of influence.

Step (ii): Identifying and ranking the most frequently used keywords and phrases using the TF-IDF function (1) [18].

$$w_{tfidf}(t_i, d_j) = \#(t_i, d_j) \cdot \log \frac{\#D}{\#(t_i, D)} \quad (1)$$

$w_{tfidf}(t_i, d_j)$ represents the quantified weight that a term t_i has over a document d_j in a collection, D is the set of all documents in the collection, $\#(t_i, d_j)$ - the term frequency in a document d_j , $\#D$ - the total number of documents in the collection, and $\#(t_i, D)$ - the document frequency, that is the number of documents in the collection in which t_i appears. This function embodies the idea that the more frequently a term occurs in a document and the fewer documents a term occurs in, then the more representative it is of that document [18]. The term t_i is selected as a meaningful word in document d_j if $w_{tfidf}(t_i, d_j) \geq \varepsilon$ where ε denotes an empirically validated threshold value.

Step (iii): Associating the most frequently used keywords and phrases with *OntoRo* and *CTA* concepts, computing the weight of the concepts $w_c(d_j)$ using (2), ranking

them accordingly and tagging the images with those concepts, which have the highest weight.

$$w_c(d_j) = \sum_{i=1}^n \left(k_{CTA} \cdot w_{tfidf}(t_i, d_j) \cdot \frac{1}{C(t_i)} \right) \quad (2)$$

Where n is the number of terms in the document that contains a concept C . In (2) w_{tfidf} , computed using (1), denotes the significance of a certain term. At this point, the weight of terms included in the *CTA* ontology is increased by a coefficient of 1.5 (see k_{CTA} in formulae (2)), to reflect their importance in the domain of interest. The value of k_{CTA} is 1 for all other terms. The use of $\frac{1}{C(t_i)}$ in (2) is based on empirical

observations and the idea that monosemic words are more domain-oriented than polysemic ones, and provide a greater amount of domain information. This converges with the common property of less frequent words being more informative, as they typically have fewer senses [19]. Therefore, the polysemy of each word influences the probability of the word to belong to a certain concept. That is because in *OntoRo* the senses of each word are coded as concept groups, which are in turn related to *OntoRo* concepts. Therefore the number of concepts that the word relates to influences the polysemy of that word. Words that relate to one concept only are therefore more significant for the domain than words that relate to more concepts. This idea is further illustrated in subsection 3.2.

Finally, concepts are ranked according to their significance $w_c(d_j)$ and those with the highest ranking are assigned to the document and its images.

3.3 Illustrative Examples

Fig. 1 shows an example, which illustrates the concept indexing algorithm developed. The first step involves building a collection. For the purposes of this example, it is assumed that the collection contains 2 million documents. Next, for each word and phrase in the collection, the algorithm determines the number of its occurrences within a document, and within the whole collection. In this example, the term t_i ‘car’ appears twice in a document d_j , and in 200 documents within the collection. Hence, $\#(t_i, d_j) = 2$, $\#D = 2,000,000$, $\#(t_i, D) = 200$, and the weight of the term ‘car’ in relation to this particular document is $w_{tfidf}(t_i, d_j) = 8$.

In step (iii), the algorithm associates all keywords and phrases significant for this page with *OntoRo* and *CTA* concepts. Significant are those keywords and phrases which have tf-idf values above a certain threshold, in this case $\varepsilon \geq 4$ based on empirical evidence. In this particular example, a second term ‘concept car’ has also been found to be of importance with a tf-idf value of 12. For the purpose of this example it is assumed that the term ‘car’ is associated with only three *OntoRo* concepts: those with numbers 222 (the main word in this concept category being ‘road’), 267 (‘travel’), and 274 (‘vehicle’). The term ‘concept car’ is related to only one *OntoRo* category, number 274 (‘vehicle’). Finally, all concepts which are found to be of importance are ranked according to

their weight and their numbers are used as tags for the images contained in those pages. The images in this example will be tagged with concept number 274 ('vehicle') as its weight is 14.66 compared to the weight of concept 222 (road) estimated as 2.66.

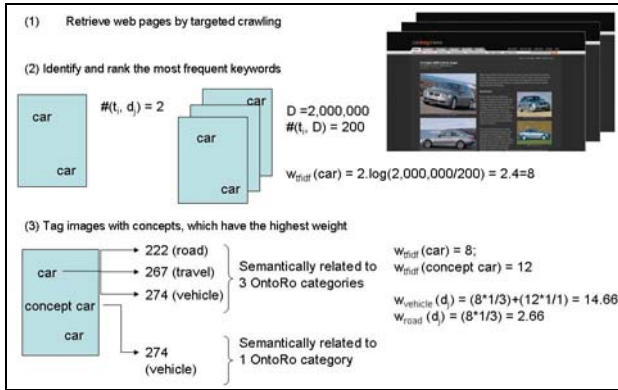


Fig. 1. Example illustrating the ontology tagging algorithm

The second example shown in Fig. 2 illustrates the benefits of using concepts when tagging images. The web page used contains the following text fragment:

Millennial Appeal Designing Concept Cars for a New Generation Automobile designers are looking beyond the Baby-Boomer generation for inspiration from the newest and most ethnically diverse group of consumers – the Millennials. This group, ages five to 24, is estimated to out number the Baby-Boomers by nearly 33 percent. With this realization, the race is on to capture the loyalty of these young consumers, who already spend between \$13 and \$27 billion annually. The auto industry has focused its attention on the youth market, as evident by the latest batch of automotive concepts. When Millennials were asked what they want in a vehicle, their answer was simple: something sporty, affordable and capable of carrying their gear and friends. Designers responded with a new group of youth-focused concepts that are compact cars, trucks and SUVs, which offer performance, cargo space, an attractive price and style reflective of this generation’s unique tastes.

As shown in Fig. 2, the most important key words and phrases in this text ('baby-boomer', 'concept car', 'generation', etc.) have been linked to concepts 126 ('newness'), 274 ('vehicle'), 45 ('union'), 110 ('period'), etc. After these concepts have been ranked, the image has been tagged with concepts 126 ('newness'), 274 ('vehicle'), 875 ('formality') and 812 ('cheapness'). As a result, a semantic search using the term 'fresh' (related to concepts 15 ('different'), 21 ('original'), 126 ('new'), etc.) would retrieve this image as relevant. A keyword-based search using the same term would naturally produce no result because the term is not contained in the text. Similarly, the use of the CTA adjectives 'trendy', 'yuppie', 'modern' in a keyword-based search would give no results. The same terms used in a semantic search would retrieve this page as 'trendy', for instance, is related to 'modernist', which is an instance of 126 ('new'). This illustrates a process called semantic query expansion, where each query is also processed semantically using the same ontological resources.

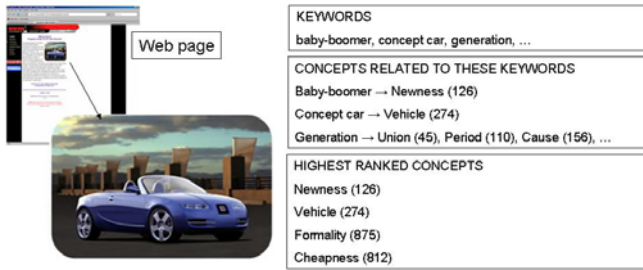


Fig. 2. Example illustrating the use of ontology tags in semantic searches

3.4 Evaluation

The evaluation of OntoTag has gone through *three stages*, described in detail in an extensive report [20]. Stage 1 included examination of the TRENDS collection and manual evaluation of 500 pages with the purpose of identifying those cases, which challenge the semantic indexing algorithms developed previously. The conclusions were used to improve the performance of the ontology tagger. The experiments showed the effort required to create a tagged corpus, and problems in the manual evaluations. Stage 2 aimed at assessing the operational capability of the ontology tagger. The results showed that 7 out of 10 pages were tagged with highly relevant concepts. The goal of stage 3 was to assess whether the semantic tags improve precision by comparing OntoTag to a standard indexing desktop piece of software (Copernic). Semantic adjectives were used to form the queries, and the results showed that in 7 out of the 10 queries conducted, OntoTag produced equally good or better results. In addition, OntoTag was evaluated through *user-centered studies*, which validated the performance of the TRENDS software by its end-users [20].

4 Conclusions

The ontology tagger OntoTag developed has demonstrated good performance and scalability, and has been integrated with a keyword-based indexing and content retrieval algorithms in the final TRENDS prototype. After extracting and ranking all significant keywords and keyphrases in a text, OntoTag extracts all related concepts and ranks them accordingly. As a result, it produces a set of concept numbers for each text, which is then used during the information retrieval. The concept indexing algorithm is currently been used as a research tool in three ongoing research projects, which explore the use of more complex ontological relationships.

Acknowledgments. The authors are grateful to their industrial collaborators Stile Bertone, Italy, Fiat, Italy, Pertimm, France and Robotiker, Spain, and the European Community for funding the research described in this paper.

References

1. Westerman, S.J., Kaur, S., Mougenot, C., Sourbe, L., Bouchard, C.: The impact of Computer-based Support on Product Designers' Search for Inspirational Materials. In: Proc. of the 3rd I*PROMS Int. Conf., Cardiff, UK, July 2-13, pp. 581–586 (2007)
2. Setchi, R., Tang, Q.: Semantic-based Representation of Content Using Concept Indexing. In: Proc. of the 3rd I*PROMS Int. Conf., Cardiff, UK, July 2-13, pp. 611–618 (2007)
3. Bouchard, C., Mougenot, C., Omhover, J.F., Setchi, R., Aoussat, A.: Building a Domain Ontology for Designers: Towards a Kansei Based Ontology. In: Proc. of the 3rd I*PROMS Int. Conf., Cardiff, UK, July 2-13, pp. 587–592 (2007)
4. Tsai, C.-F.: A Review of Image Retrieval Methods for Digital Cultural Heritage Resources. *Online Information Review* 31(2), 185–198 (2007)
5. Ferecatu, M., Boujemaa, N., Crucianu, M.: Semantic Interactive Image Retrieval Combining Visual and Conceptual Content Description. *ACM Multimedia Systems Journal* 13(5-6), 309–322 (2008)
6. Enser, P.G.B., Sandom, C.J., Hare, J.S., Lewis, P.H.: Facing the Reality of Semantic Image Retrieval. *Journal of Documentation* 63(4), 465–481 (2007)
7. Setchi, R., Tang, Q.: Concept Indexing Using Ontology and Supervised Machine Learning. *Transactions on Engineering, Computing and Technology* 19, 221–226 (2007)
8. Köhler, J., Philippi, S., Specht, M., Rüegg, A.: Ontology Based Text Indexing and Querying for the Semantic Web. *Knowledge-Based Systems* 19(8), 744–754 (2006)
9. Conesa, J., Storey, V.C., Sugumaran, V.: Improving Web-Query Processing Through Semantic Knowledge. *Data and Knowledge Engineering* 66, 18–34 (2008)
10. Gene-Ontology-Consortium. Creating the Gene Ontology Resource: Design and Implementation. *Genome Res.* 11(8), 1425–1433 (2001)
11. WordNet, <http://wordnet.princeton.edu/> (last accessed 24/03/2009)
12. Cyc, <http://www.ontotext.com/downloads/CycMDB.html> (last accessed 24/03/2009)
13. Kiryakov, A., Popov, B., Terziev, I., Manov, D., Ognyanoff, D.: Semantic Annotation, Indexing, 1st Retrieval, Web Semantics. *Science, Services and Agents on the World Wide Web* 2(1), 49–79 (2004)
14. Dill, S., Eiron, N., Gibson, D., Gruhl, D., Guha, R., Jhingran, A., Kanungo, K., McCurley, S., Rajagopalan, S., Tomkins, A., Tomlin, J.A., Zien, J.Y.: A Case for Automated Large-Scale Semantic Annotation, Web Semantics. *Science, Services and Agents on the World Wide Web* 1(1), 115–132 (2003)
15. Corcho, O.: Ontology Based Document Annotation: Trends and Open Research Problems. *International Journal of Metadata, Semantics and Ontologies* 1(1), 47–57 (2006)
16. Benjamins, V.R., Contreras, J., Blazquez, M., Doderio, J.M., Garcia, A., Navas, E., Hernandez, F., Wert, C.: Cultural Heritage and the Semantic Web. In: Bussler, C.J., Davies, J., Fensel, D., Studer, R. (eds.) *ESWS 2004*. LNCS, vol. 3053, pp. 433–444. Springer, Heidelberg (2004)
17. Davidson, E. (ed.): *Roget's Thesaurus of English Words and Phrases*. Penguin, UK (2003)
18. Salton, G., Buckley, C.: Improving Retrieval Performance by Relevance Feedback. *Journal of the American Society for Information Science* 41(4), 288–297 (1990)
19. Gliozzo, A., Strapparava, C., Dagan, I.: Unsupervised and Supervised Exploitation of Semantic Domains in Lexical Disambiguation. *Computer Speech and Language* 18(3), 275–299 (2004)
20. Ontology Tagger (June 12, 2008), <http://www.trendsproject.org> (last accessed 25/03/2009)

Design and Implementation of a Methodology for Identifying Website Keyobjects

Luis E. Dujovne and Juan D. Velásquez

Departamento de Ingeniería Industrial, Universidad de Chile
ldujovne@dcc.uchile.cl, jvelasqu@dii.uchile.cl
<http://wi.dii.uchile.cl>

Abstract. Rich media websites like Flickr or Youtube have attracted the largest user bases in the last years, this trend shows web users are particularly interested in multimedia presentation formats. On the other hand, Web Usage and Content Mining have focused mainly in text-based content. In this paper we introduce a methodology for discovering Website Keyobjects based in both Web Usage and Content Mining. Keyobjects could be any text, image or video present in a web page, that are the most appealing objects to users. The methodology was tested over the corporate site of dMapas a Chilean Geographical Information Systems service provider.

Keywords: Web Mining, Website Keyobjects, Web User Preferences.

1 Introduction

Webmasters are always looking for new ways to enhance the content of their site in order to attract and retain users. If they can gain knowledge of their user preferences, they could offer the content users are looking for. In [1] a methodology for discovering website keywords is proposed giving clues of what content users are looking for.

In this paper, we propose a generalization of this methodology which allows the discovery of the objects that attracts the attention of most users. This objects can be not only text, but also images, videos, etc.

This paper is organized as follows. Section 2 provides an overview on related work. In section 3 we describe the methodology used to find Website Keyobjects. Section 4 describes the application of our work in an actual website and Section 5 concludes this work and points out future work.

2 Related Work

The methodology to extract Website Keyobjects must understand what a web object is and which are most appealing to users. To obtain this, IR techniques will be used along with web usage mining to infer the user preferences. To define web objects, special metadata will be used.

2.1 Metadata

In order to define web objects, there must be a way to include computer retrievable information about the content of resources whose format doesn't allow this, such as images or videos. This can be made by providing metadata; this is "data about data".

There are many different ways in which metadata can be incorporated to a website, from models proposed by the W3C in the Resource Description Framework (RDF) [2] or by the Dublin Core Initiative [3], to others adjusted from metadata models that haven't been made especially for the web. RDF and Dublin Core are XML based metadata models which describe resources in the web. RDF is based on triplets of subject, predicate and object. On the other hand, Dublin Core is comprised of a series of standards that using several basic terms arranged in different levels defines an ontology that describe a resource.

Another metadata model worth describing, was developed by the Moving Pictures Expert Group (MPEG) [4] to add metadata to videos. In this model, the MPEG defines several descriptors in XML, which allows adding information to videos, ranging from purely technical issues, to information about its content. Then, every frame of a video is related to their corresponding XML files allowing complex queries to be ran over them. Using an approach similar to the MPEG metadata description scheme, web objects can be described.

2.2 A Methodology to Discover Website Keywords

The basis of this work is the methodology created by Velásquez et al [1] for identifying website keywords. In this work, IR and WUM techniques were used to find the keywords most appealing for the users of a website.

The methodology begins by applying the Vector Space Model (VSM) [5] to the web data, so given R words found in Q documents, a weight matrix of dimension $R \times Q$ is created

$$M = (m_{ij}); i = 1 \dots R, j = 1 \dots Q \quad (1)$$

where m_{ij} is the weight of word i in document j .

This weight is defined using the TF*TFIDF (Term Frequency Times Inverse Document Frequency) [5]. This weight considers the fact that some words are more relevant than others, and its calculated based in the term frequency, which is the number of times a word i appear in a given page j . The other term used to calculated the word is the inverse document frequency, where the number of documents in which word i appears, n_i , is calculated inversely with respect to the whole set Q , this is $\log(\frac{Q}{n_i})$. Once these weights have been calculated, it is possible to establish a measure of distance between two documents by calculating the cosine distance between two vectors as shown below.

$$pd(p_i, p_j) = \frac{\sum_{k=1}^R m_{ki} m_{kj}}{\sqrt{\sum_{i=1}^R (m_{ki})^2} \sqrt{\sum_{i=1}^R (m_{kj})^2}} \quad (2)$$

This vectors represents a given page an corresponds to a column from matrix M in equation (1).After this is calculated for every document, the next step is to retrieve the user sessions from the website log. This log contains information that enables the recreation of user sessions by correctly filtering and sequencing the information present in the log [6] in a process named Sessionization.

After user sessions have been reconstructed, a vector named "User Behaviour Vector" (UBV) is created, this vector considers the i-most important pages for each user session and is represented as follows.

$$v = [(p_1, t_1) \dots (p_n, t_n)] \tag{3}$$

This vector has i-tuples in the form (p,t) where p is the index representing a page in the website, and t is the percentage of time spent in that page in relationship with the whole session, this gives a short and precise description of a user session.

As users session have to be clustered, the similarity measure in equation (4) is used.

$$st(\vartheta_i(\alpha), \vartheta_i(\beta)) = \frac{1}{i} \sum_{i=1}^n \min \left\{ \frac{\tau_k^\alpha}{\tau_k^\beta}, \frac{\tau_k^\beta}{\tau_k^\alpha} \right\} * dp(p_k^\alpha, p_k^\beta) \tag{4}$$

where ϑ is the session of users α and β , and $\min\{\cdot, \cdot\}$ is the ratio between the percentage of time spent and $dp(\cdot, \cdot)$ is the similarity measure between pages (2).

This similarity measure is used in the clustering process from which the website keyword can be retrieved using equation (5).

$$kw[i] = \sqrt{\prod_{p \in \varsigma} m_{ip}} \tag{5}$$

Where kw is an array that contains the weights for each word related to a given cluster, and ς is the set of all documents that belong to that cluster. To retrieve the keywords, the array is ordered from high to low so the first elements are considered to be the keywords.

3 A Methodology to Discover Website Keyobjects

Based on the methodology created to discover website keywords, it is possible to create a more general model that will enable the discovery of not only keywords, but keyobjects. In this section we will show this methodology along with the definition of a web object and the tools used to describe them.

3.1 Web Objects

Websites are composed primarily of free text, but they also have other data formats such as images, videos, etc. The later formats do not provide information about their content in a way that can be easily retrieved by computers, so little or no content analysis can be made over them.

To attend this issue we introduce web objects, which are defined as "any structured group of words or multimedia resource within a web page that has

metadata to describe its content” and Website Keyobjects are defined as “The web objects or groups of web objects that attracts web users attention and that characterize the content of a given webpage or website.”

The implementation of web objects can be made in several ways, because it relies heavily on the ontology used to describe them. In this work we created a simple ontology where an XML document describing every object in a particular page is created. To establish a link between an object in the page and the XML document, standard HTML tags are used.

In the XML document, each object is characterized by an identifier, its format and a list of concepts that describe its content. Each concept is a group of three substantives, which in Spanish language allows a *sufficient* but not *complete* definition of any concept [7]. Also every concept belongs to a certain category that groups concepts together. By using these categories the concepts that describe web objects are able to relate with each other.

To compare objects we use the fact that a web object is really a bag of concepts that defines its content, so given two objects every concept between them are compared. This comparison is made word by word using a thesauri. The most similar with each other are matched.

Once all the concepts are matched, they are sorted in a way such, that every concept is in the same relative position within each object. Then a string representing the object, that has a symbol representing each concept category is created. This string has the following structure.

$$O = Category_1, \dots, Category_N \quad (6)$$

If the objects are characterized in this way and for each category a different symbol is used, the two strings that represent each object can be compared using the *Levenshtein* Distance [8], also known as Edit Distance. Finally the distance between two objects is calculated using equation (7).

$$do(O_1, O_2) = 1 - \frac{L(O_1, O_2)}{\max(|O_1|, |O_2|)} \quad (7)$$

Where $L(\cdot, \cdot)$ is the Levenshtein distance between two objects depicted as the Strings in equation (6).

3.2 Sessionization and User Behaviour

The methodology uses the session reconstruction method proposed in [6], which result in the grouping of every page within a user session. As a web page is composed by one or more objects, each request in every session is *“expanded”*, so that for each request for a page, the objects it contains are considered as an independent request. To determine the time spent by a user in each object, an approximation is used, constructed by taking a survey over a controlled group of users asking which are the most appealing objects to them, and grading them accordingly.

The user behaviour is obtained in [1] using the vector in equation (3). In this work this is defined similarly by replacing pages with objects as shown in equation (8) and it is named *Important Object Vector* (IOV).

$$v = [(o_1, t_1) \dots (o_n, t_n)] \tag{8}$$

The similarity between this vector is also defined similarly to (4) resulting in the following equation.

$$st(\vartheta_i(\alpha), \vartheta_i(\beta)) = \frac{1}{i} \sum_{k=1}^n \min \left\{ \frac{\tau_k^\alpha}{\tau_k^\beta}, \frac{\tau_k^\beta}{\tau_k^\alpha} \right\} * do(o_k^\alpha, o_k^\beta) \tag{9}$$

where $do(\cdot, \cdot)$ is defined in (7).

3.3 User Sessions Clustering

Different clustering techniques can be applied to group user session, in this work *Kohonen's Self Organizing Feature Maps* (SOFM) [9] was used, and cross-checked with *K-Means*, and *Association Rules*.

SOFM is a special type of neural network, where typically a two-dimensional grid of neurons is ordered, so it reflects changes made in the n-dimensional vector the neurons represent. In this case, these vectors are IOV's. SOFM works with the concept of neighbourhoods, and the way in which these are set-up defines the topology of the network. In the case of this work, a two-dimensional toroidal topology will be used, where each neuron will be neighbour of their up-down, left-right neurons. The map has four edges, so in the case of the upper left edge neuron its neighbours will be their left and down neurons and their opposite upper and left neurones. For the other edge neurons, their neighbourhood is defined analogously. A figure of this network can be found in [1].

The update rule for SOFM is separated between one for time, and another for objects, the first is given in equation (10) and the second in equation (11)

$$\tau_{N_i}(t + 1) = \tau_{N_i}(t) + \frac{1}{\sqrt{t}} e^{-\frac{1}{2t}} (\tau_{E_i} - \tau_{N_i}) \tag{10}$$

where τ is time a user spent in a particular page in a session described as in (8), t is the epoch, N_i is the BMU and E_i is the example presented to the network.

$$o_{i+1}^N = \gamma \in \Gamma / D'_{NE} \approx do(\gamma, o_i^N) \tag{11}$$

where $\Gamma = \{\gamma_1, \dots, \gamma_n\}$ is the set of all objects in the site, and DNE is defined by the equation:

$$D_{NE} = [do(o_1^N, o_1^E), \dots, do(o_n^N, o_n^E)] \tag{12}$$

4 Practical Application

The methodology was applied over the corporate website a Chilean geographical information systems service provider which had 27 static pages, completely written in Spanish, presenting content in free text, images and flash animations. The weblog used corresponds to the month of June 2007 and has 31.756 requests.

The webmaster identified 40 objects, 26 are text-based, 11 are images and 3 are flash animations. 344 concepts were associated to these objects and these concepts are categorized in one of the 12 categories. Depending on the context in which an object is positioned, two concepts can belong to different categories even if they are exactly the same.

4.1 Similarity between Objects

An important part of the work is the algorithm to compare two objects. A proof of concept was performed before the clustering process began proving the similarity measure created was suitable enough to compare conceptually two objects.

This proof considered a dataset of four objects, two were flash animations depicting demos of the GIS solutions dMapas provides, the other two were free text within tables, the first describes from a technical point of view GIS Systems, while the other shows information about the company, their owners and employees. The algorithm was performed over these objects. The results were checked and approved by the webmaster, showing the objects are correctly compared.

4.2 Results

The sessionization process was implemented in **PHP** using a reactive strategy [6] and considering a limit of 30 minutes per session [10], resulting in 5.866 sessions over 19.282 requests giving an average of 3.29 objects per session. To create the IOV the number of objects considered was calculated by taking the mean number of objects per session and adding the standard deviation which led to all sessions having 6 or more requests. Using this constraint, only 1.464 out of the 5.866 sessions were used.

SOFM was implemented using Python, K-Means using Java and Association Rules using Weka [11], all of them ran over an Intel T2300 Core Duo running at 1.63GHz with 1Gb in RAM with Windows XP. SOFM algorithm ran in 2 hours while K-means in 15 minutes and Association Rules in approximately 30 minutes.

The objects were labeled according to their main topic and a cluster is accepted only if the objects it contains have no more than two topics. With the help of experts, 9 clusters were identified for SOFM, 5 for K-Means and 7 for Association Rules. These are shown in table 1.

For determining the website keyobjects, the objects in the centroid of every cluster were counted for every algorithm. The ones that appeared most frequently are considered the website keyobjects and are shown in table 2.

It can be seen that from the top 10 objects, that 7 are presented in text format, 2 as flash animations and only one corresponds to an image. They focus on a small section of the company's site leaving out a lot of information that administrators assumed to be very useful to users.

These keyobjects were cross-checked with the results of the survey used to determine the time spent in an object leading to a matching of 87% between the objects discovered by the methodology and the ones chosen by users.

Table 1. Clusters Discovered using SOFM, K-Means and Association Rules

SOFM	K-Means	Association Rules
Cartography	Cartography	Cartography
Geobusiness	Geobusiness	Geobusiness
Demos	Cartography and GIS	Cartography and GIS
Geobusiness and GIS	Geobusiness and GIS	Geobusiness and GIS
The Company and Cartography	Demos and Cartography	Cartography and The Company
Demos and Cartography		Cartography and Geobusiness
Demos and GIS		Demos and Geobusiness
The Company and GIS		
Cartography and GIS		

Table 2. Website Keyobjects

Object	Type	Concept	Count
Index	Flash	General Information about dMapas Products	28
Cartography 1	Text	Technical Information about Cartography	23
GIS Products 1	Text	General Information about GIS Products	17
GIS Products 2	Text	General Information about GIS Products	14
Cartography 2	Text	Technical Information about Cartography	14
Cartography Products	Text	Information about Cartography Provided by dMapas	13
About GIS	Text	General Information about GIS Systems	10
Demo 2	Flash	Demonstratin of a GIS Application	10
Geobusiness	Image	Information about Geobusiness	9
Cartography 3	Text	Technical Information about Cartography	9

5 Conclusions

In this work a methodology for discovering website keyobjects was introduced. Web objects are defined as *any structured group of words or multimedia resource within a web page that has metadata to describe it's content*, and website keyobjects are the web objects that most attracts the users interest.

The methodology uses clustering algorithms to group web user sessions together. To implement web objects, an ontology was created to add metadata to the site and a similarity measure used to compare two web objects was introduced. These are proved to be valid by a comparison of the produced results with the experts opinion.

The website keyobjects discovered for the site of the Chilean GIS service provider dMapas website were checked with a selected group of users, proving it correctness and can be used to enhance the content of the site pointing out not only the information users are looking for but also, which presentation formats are the most appealing to them.

As future work, an improvement of the ontology used can be done, so it allows more complex and complete definitions of objects, with this new definition the similarity measure can also be improved. Finally, work can be done in finding a way to determine the time spent by a user on a certain object within a page in

an automatic or semi-automatic way, as opposed to the construction algorithm used in this work based on the results of a survey took to a selected group of users, asking which objects are the most appealing to them.

Acknowledgement

This work has been partially supported by the Chilean Millennium Scientific Institute of Complex Engineering Systems.

References

1. Velásquez, J.D., Bassi, A., Yasuda, H., Aoki, T.: Towards the identification of keywords in the web site text context: A methodological approach. *Journal of web information systems* 1, 11–15 (2005)
2. W3C Semantic Web Group, Resource Description Framework Specification: Concepts and Abstract Syntax (2004)
3. Dublin Core Metadata Initiative (DCMI), <http://dublincore.org/>
4. Salembier, P., Smith, J.: MPEG-7 multimedia description schemes. *IEEE Transactions on Circuits and Systems for Video Technology* 11(6), 748–759 (2001)
5. Baeza-Yates, R., Ribeiro-Neto, B.: *Modern Information Retrieval*. Addison-Wesley, Reading (1999)
6. Berendt, B., Spiliopoulou, M.: Analysis of navigation behaviour in web sites integrating multiple information systems. *The VLDB Journal* 9, 56–75 (2000)
7. Seco, M.: *Problemas Formales de la Definición Lexicográfica*. In: *Estudios de Lexicografía Española*, Gredos, Madrid (2003)
8. Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions and reversals. *Sov. Phys. Dokl.*, 705–710 (1966)
9. Kohonen, T.: *Self-Organizing Maps*. Springer, Heidelberg (2001)
10. Catledge, L., Pitkow, J.: Characterizing browsing behaviours on the world wide web. *Computer Networks and ISDN systems* 26, 1065–1073 (1995)
11. Weka Open Source Machine Learning Software, <http://www.cs.waikato.ac.nz/ml/weka/>

NLP Contribution to the Semantic Web: Linking the Term to the Concept

Gaëlle Lortal¹, Nathalie Chaignaud², Jean-Philippe Kotowicz²,
and Jean-Pierre Pécuchet²

¹ THALES

`gaelle.lortal@thalesgroup.com`

² INSA Rouen - LITIS EA 4108

`{chaignaud,kotowicz,pecuchet}@insa-rouen.fr`

Abstract. The Semantic Web (SW) originally aims at studying a system interoperability based on a shared common knowledge base (ontology). Henceforth, the SW sets its heart on a semantic coordination of community parlance representative resources (in complement to a common knowledge base shared by the users). The matter is not only to use techniques to handle a large amount of data, but also to use approaches to keep the community parlance features. Thus, Web documents and folksonomies are the main semantic vehicle. They are little structured and Natural Language Processing (NLP) methods are then beneficial to analyze language specificities with a view to automating tasks about text. This paper describes a use of NLP techniques for the SW through a document engineering application: the information retrieval in a catalogue of online medical resources. Our approach emphasizes benefits of NLP techniques to handle multi-granular terminological resources.

1 Introduction

Natural Language Processing (NLP) can be used into two different ways that change its relation with the Web. It is either a means to guide the linguists in their analysis and then, it uses the Web as corpus, or it is a tool used in the design of applications insofar as it implements linguistic theories. These latter aspects of NLP interest particularly the Semantic Web (SW), though, NLP and SW are not really linked. “Using the Web as a corpus” supposes the next definition: “A corpus is a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language” ([1] p. 4). This definition represents the descriptive and interpretative process in the corpus linguistics as well as its quantitative interest. According to [2], corpus linguistics aims at constituting and giving availability to resources for their comparison by linguists. Indeed, [3] distinguishes corpus-based from corpus-driven studies. Corpus-driven studies use a corpus producing “meaning” (the corpus is not only a support, but also a contribution to the theory and it fundamentally helps to build the theory). Corpus-based studies use a corpus as an a-priori theory recollection (the corpus has then a validating or picturing

function and reveals a pre-existing meaning). We consider the Web as a corpus to carry on the social phenomena to the linguistic phenomena and then we use a corpus of Web Medical Library to improve NLP based document retrieval application. The design of NLP applications is also based on principles upon the language. The first approach is statistical, considering text/documents as bags of words and enabling to fast process large amount of data. The second approach is the symbolic approach which considers the corpus as subtly parsed according to several layers of linguistic analysis (morphological, lexical, syntactical, semantic, pragmatic...). Nowadays, the main contribution of NLP to SW consists in the syntactic processing of the text at the document/system interface. The Web Medical library we use has already brought up the question of how the users can access the data they need when they do not know the classification in use. Actually, the users need to use their own keywords as well as the system requires to use its own classification. Given that the semantic management of the document is limited, we consider that NLP can bring some help to SW at the human/system interface in order to contextualize or to profile the user. The human/system interface involves not only to be able to handle large data by statistical means, but also to fit the users classification to match their search needs. Both approaches have benefits and drawbacks and then we propose to use a combination of these two approaches to improve the semantic layer analysis.

More precisely, the solution developed here deals with medical document search led by a matching between user concept definitions and system concept definitions. This application lays the emphasis on the necessity of both a general terminology and a specialized terminology to obtain relevant documents.

This paper is organized as follows: Section 2 defines what a term is and what it represents in a classification used to index/retrieve documents for a specific user. Moreover, we present the basics in NLP methods for Information Extraction (Syntactic tagging, terms and relations extraction). In Section 3, we unfold our problem of specialized indexing for non-specialized search on a Web medical library. Then, we explain, in Section 4, our terminology alignment by means of various NLP methods. Their combination enables to deal with the complexities of the Web documents (amount, variability, ...).

2 Related Work

Nowadays, Web documents are available thanks to annotations linked to a classification used by the search engines. Some tools propose NLP modules to automatically or semi-automatically process Web documents for information retrieval or indexing tasks ([4], [5]). But to be efficient, this classification has to be well suited not only to the system but also to the user. To this aim, a matching of the user's classification with the classification of the system can be proposed. NLP can be used to automatic or semi-automatic process Web documents for information retrieval or indexing tasks but also for adaptation to the user. As our aim is more to fit a user classification to a common classification than to build an adapted classification to the user, we first position our work with regard

to the notion of term in indexation and its use in a community, and then we present NLP methods and tools for information extraction based on a corpus.

2.1 Terminology and Communities

Terminology is the study of terms that denominate classes of objects and concepts. Terms (signifier) are words or expressions having a specific meaning in a particular context and represent the linguistic aid of the concepts (signified). The link between the term and the concept is preferably biunivocal, that is to say that the term must have only one meaning. Terms chosen as concepts in a domain allow to index resources.

Moreover, a term denotes the belonging to a community. In a particular professional domain, speakers using the same vocabulary (technical or specialized language) constitute a linguistic community. The virtual Web communities in a particular domain use a specific vocabulary called “sociolect”. Members of such a community build together a non-systematic and a non-formal classification. On the contrary, classifications used in documentation are formal. Some researchers speak about “ethnoclassifications” and others use “folksonomies”. These semi-formal classifications are numerous on the Web and they structure it semantically. They have to be considered in order to design a huge semantic network improving Web information retrieval.

E. Wüster was the first researcher who denominates objects of the technical terminology. He dreamed of a consensual language based on an agreement of all the specialists. In the medical domain, this terminology exists as a thesaurus built by experts of the domain named MeSH (Medical Subject Headings – a controlled vocabulary proposed by the National Library of Medicine in 1960).

2.2 NLP Methods for Information Extraction

To build a corpus based terminology, the first task is to identify terms and relations between these terms. According to the linguistic approaches, term extraction uses generally tagged corpus and linguistic skeleton.

Corpus Tagging. The most famous taggers are the Brill’s tagger [6] and Tree-Tagger [7]. TreeTagger uses binary decision trees evaluating grammatical tags. The Brill’s tagger learns tagging rules from a manually annotated corpus. But tags are predetermined and are not always well adapted to the specialized texts.

Term Extraction. The use of regular expressions for linguistic skeleton is not efficient when a term appears in different forms and is linked to a context. Finite state automata (to extract words or expressions from the text), transducers (to extract morpho-lexical information or parts of speech), recursive transition networks (to transform text, e.g. negations, passive forms, etc.) and context-free grammars (to associate inflections and derivations to a word) can also be used.

Term extraction tools can be divided into three classes: statistical, linguistic and mixed approaches.

Statistical term extractors like Xtract [8] or LIKES [9] choose terms-candidate among the most frequent sequences of words, according to the hypothesis that specific terms to a domain are more frequent in specific texts. Because these methods do not need linguistic information, they can be applied easily to a new domain. However, they need a learning phase on huge corpora and their results present some noise.

Tools using linguistic approach are based on various linguistic analysis, for example, a syntactic analysis allows to extract noun phrases as relevant terms-candidate (e.g. Lexter [10] and XIP [11]).

Both statistical and linguistic methods are interesting, thus mixed approaches have been proposed (ACABIT [12] or Exit [13]). They select a list of terms-candidate by calculating frequencies and then they refine their choices by means of linguistic skeleton.

Relation Extraction. The aim of the task is to identify relations between or among terms in order to build semantic classes. Only linguistic approaches are described here.

Zellig [14] is based on the hypothesis that syntactic regularities in the term context allow to build semantic classes. However, the classes found by Zellig are often too wide even if they can be refined afterwards.

Cameleon [15] is an interactive tool that helps the user to find semantic relations by means of lexico-syntactic skeletons. Relation items as verbs are considered specific to the corpus.

Upéry [16] allows to compare couples of terms found in the same syntactic context by using distributional proximity measures. Results are helpful to design specific ontologies.

Thus, NLP proposes interesting methods to assist the user in improving ethnoclassifications for information retrieval.

3 A Specialized Indexing for a Non-specialized Search

The project VODEL (french acronym for Ontological Valorization for Electronic Dictionaries) aims at allowing a non-expert user to search documents indexed with an expert terminology (or an ontology). That means that our main goal is to link general vocabulary (the one of the user's query) with specialized vocabulary (the one of the search engine).

Our application domain is the Web system CISMef (french acronym for "Catalog and Index of French language health resources", www.cismef.org). CISMef has to assist health professionals and patients in their document search. CISMef is based on the thesaurus MeSH to index documents. To this end, it uses four different concepts: *meta-term* (112 in number) corresponding to medical or biological specialized areas, *keyword* (12369) used to define the topics handled, *subheading* (83) making the meaning of a keyword precise and *resource type* (280) describing the nature of the document.

The problem comes from the fact that patients use everyday words and they do not know and they do not understand the medical vocabulary used by the

system. Our goal is to bring some help to the user. Nowadays, CISMef uses a conventional algorithm to search documents. We have to propose an algorithm based on knowledge extraction methods using a deeper linguistic analysis to link general vocabulary with specialized vocabulary.

For this work, only French has been considered.

Thus, the specialized terms comes from the CISMef terminology. However, we also use the “VIDAL de la famille” (for families), which is a French medical dictionary. The definitions are given in an understandable way for common people but only about medicine.

The general knowledge (semantic and syntactic) comes from the tools of LDI (www.dictionnaire.sensagent.com) and Wikipedia (www.wikipedia.org). LDI [17] is a semantic network: an oriented graph with three semantic models (including WordNet - wordnet.princeton.edu) that are complementary. The relations connect “synsets” but also concepts. All the relations (around 100 types) are typed and balanced. Comparison between words are done using the LDI’s structure and the proximity measure of common subsumers. Words and their environment are used as starting point for definition alignment.

Here is an example of the contribution of general dictionary to document search in CISMef: when a patient searches documents about the french expression “médicaments pour mal de crâne”, the colloquial expression for “drug for headache”, no results are found. However, LDI proposes the next definitions for “mal de crâne”:

```
névrалgie[Class] (in English, neuralgia)
mal de tête[Theme] (in English, headache)
céphalée, migraine[Spec.] (in English, cephalgia, migraine)
```

Then, by semantically linking “medicine” with “therapy” and by using the terms “cephalgia” and “therapy” in the query, CISMef finds 10 relevant documents as result. That satisfies the user.

4 Terminology Alignment by Means of NLP Methods

In a first study, we run the LSA (Latent Semantic Analysis) algorithm [18] on our data but the results were poor (that is not surprising): the algorithm succeeded to align definitions using word bundles but it found too many alignments creating noise. We don’t present further result about this test here.

In this section, we propose two approaches to align different terminologies for document search: linguistic and mixed approaches. Before presenting these two methods, we describe our corpus on which processes were performed.

4.1 Pre-treatments of Our Corpus

Our corpus contains manually aligned definitions taken from four dictionaries: LDI vs. MeSH vs. Vidal vs. Wiki. In this case, we have 55 definitions for each dictionary. For each term, the definitions of the four dictionaries are aggregated.

Thus, the corpus presents four layers of specialization for each term, from the most specialized to the less specialized. Moreover, some pre-treatments were performed on these data and we kept only terms like noun (N), adjective (A) or verb (V) (for process speed reasons). The pre-treatments were performed automatically by the “Sémiographe” (a tool of LDI):

- segmentation to cut sentences into words;
- morphological process comprising the stemming (words having the same root) and the lemmatization (words having the same meaning);
- syntactic analysis for example to recognize fixed expressions;
- semantic analysis to solve problems of polysemy.

Because pre-treatments were performed automatically, the corpus presents some weaknesses: synonyms and hyponyms are not always relevant due to contextual meaning and morpho-syntactic labelling is sometimes inaccurate.

4.2 Linguistic Approach

This approach allows to link terms in a text with keywords of the terminology. To this aim, we used transducers for their flexibility. They are easy to design and well-adapted for alignment problems but they need a precise linguistic analysis of the corpus. For our study, contributions of transducers are twofold: on the one hand, they help to extract terms for indexing and on the other hand, they allow to link general terms with specialized terms.

[19] presents transducers for the CISMeF application in order to develop automatic indexing or controlled indexing with Nooj [20]. We have refined these transducers to link general expressions with keywords or subheadings of the CISMeF terminology. The methodology we used consists in: interviewing the expert to bring out some hints to link general expressions with keywords or subheadings; identifying recurrent forms in the corpus and translating them as general rules; validating all the transducers by the expert before implementation.

Transducers are manually built, so their design is very expensive. There exists one transducer per subheading (83).

4.3 Mixed Approach

A mixed approach allows to link the text and its topics with keywords. Our work is based on statistical clustering improved by linguistic features [21]. After a learning phase, the clustering algorithm can link each word from a definition corpus with its definitions in order to propose topics of a document.

A first validation has been done with classical statistical methods (Spearman, Dice, Jaccard, Levenshtein, etc.) to link definition of CISMeF concepts with general definition (from LDI) of these concepts. But the results were not good (F-Mesure = 63%).

Another method has been set up to align terminologies using Support Vector Machine (SVM) [22] classifier with a Radial Basis Function (RBF) kernel.

Table 1. F-Measure and confidence interval for terminology alignment

	<i>LDI vs MeSH</i>	<i>LDI vs Vidal</i>	<i>LDI vs Wiki</i>	<i>MeSH vs Vidal</i>	<i>MeSH vs Wiki</i>	<i>Vidal vs Wiki</i>	<i>LDI, MeSH, Vidal vs Wiki</i>
<i>N</i>	81 ± 3.15	82 ± 3.09	84 ± 2.95	81 ± 3.15	80 ± 3.21	87 ± 2.70	98 ± 0.69
<i>NA</i>	77 ± 3.38	74 ± 3.52	86 ± 2.79	80 ± 3.21	85 ± 2.87	88 ± 2.61	98 ± 0.69
<i>NAV</i>	77 ± 3.38	79 ± 3.27	85 ± 2.87	80 ± 3.21	80 ± 3.21	84 ± 2.95	98 ± 0.69

The corpus was divided into two subsets: a training set and a testing set. The training set was again divided into a learning and a validation sets. The learning and validation subsets were used by the SVM with parallel grid search for learning the optimal values for the parameters by using a 10-fold cross-validation. The testing set and the optimal parameters were used for classifying the instances from the test set.

Moreover, each definition was considered at a lexical level, when two words are equal if and only if they are represented by the same strings of characters. For this level we investigated three models in order to represent a definition:

- a model that takes into account the nouns only (N);
- a model that takes into account the nouns and adjectives only (NA);
- a model that takes into account the nouns, adjectives and verbs only (NAV).

Thus, the SVM classifier was adapted to our problem of definition alignment and results were excellent (F-Mesure 98%). This model takes advantage of both distance for alignment and a precise representation of the text.

5 Conclusion and Perspectives

In the project VODEL, it is essential that a question from a patient could be translated into a query using the terminology of the system, that is to say the domain terminology. To this aim,

1. we have used the term definitions in the patient’s question to build the query with specialized terms and a learning phase of the cotext;
2. transducers have been built to verify if terms of the questions are or are not hints to find keywords;
3. and word extension have been used for each word of the question linked with terms of the terminology.

These three processes improve information retrieval thanks to the advantages of both statistical and linguistic methods.

However, theoretical problems stay: once the alignment is done, which documents have to be proposed? For whom and for what use? In the patient’s question, how to take account of the user’s profile and the user intention?

Some answers are given in [23]. The question can be complex but its formulation can facilitate the access to the relevant documents. Generally the users do not contextualize their questions and do not express entirely what they want. The system has to adapt itself to answer the question as precisely as possible.

References

1. Sinclair, J.: Preliminary recommendations on text typology. In: EAGLES (Expert Advisory Group on Language Engineering Standards) (1996)
2. Pery-Woodley, M.: Discours, corpus, traitement automatiques. In: Condamines, A. (ed.) *Sémantique et corpus*, Hermès, Londres (2005)
3. Williams, G.: A corpus-driven analysis of collocational resonance in French and English Texts. Hédiard M. *Lezioni di Dottorato*, Edizioni Spartaco (2005)
4. Handschuh, S., Staab, S.: Annotating of the shallow and the deep web. In: Handschuh, S., Staab, S. (eds.) *Annotation for the semantic web*, pp. 25–45. IOS Press, Amsterdam (2003)
5. Dingli, A.: Next generation annotation interfaces for adaptive information extraction. In: 6th Annual Computer Linguists UK Colloquium, Edinburgh, UK (2003)
6. Brill, E.: A simple rule-based part of speech tagger. In: Conference on Applied Natural Language Processing, Trento, Italia, pp. 152–155. ACL (1992)
7. Schmid, H.: Probabilistic part-of-speech tagging using decision trees. In: International Conference on New methods in Language Processing, UK (1994)
8. Smadja, F.: Retrieving collocations from text: Xtract. *Computational Linguistics* 19(1), 143–177 (1993)
9. Rousselot, F., Montessuit, N.: La station de travail likes. In: INTEX Workshop (2003)
10. Bourigault, D.: Lexter, a natural language processing tool for terminology extraction. In: EURALEX International Congress, Goteborg, Nederland, pp. 771–779 (1996)
11. Ait-Mokhtar, S., Chanod, J.P., Roux, C.: A multi-input dependency parser. In: International Workshop on Parsing Technologies, Beijing, China, pp. 201–204 (2001)
12. Daille, B.: Conceptual structuring through term variations. In: ACL Workshop on Multiword Expressions: Analysis, Acquisition and Treatment, pp. 9–16 (2003)
13. Roche, M., Heitz, T., Matte-Tailliez, O., Kodratoff, Y.: Exit: Un système itératif pour l'extraction de la terminologie du domaine à partir de corpus spécialisés. In: *Journées d'analyse statistique des données textuelles*, pp. 946–956 (2004)
14. Habert, B., Fabre, C.: Elementary dependency trees for identifying corpus-specific semantic classes. *Computer and the Humanities* 33(3), 207–219 (1999)
15. Séguéla, P., Aussenac-Gilles, N.: Extraction de relations sémantiques entre termes et enrichissement de modèles du domaine. In: IC, Paris, France, pp. 79–88 (1999)
16. Bourigault, D.: Upery: un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus. In: TALN, Nancy, France, pp. 75–84 (2002)
17. Dutoit, D., Papadima, O.: Alexandria as a result of the integration of wordnet and ldi. In: International WordNet Conference, pp. 157–163 (2005)
18. Landauer, T., Foltz, P., Laham, D.: An introduction to latent semantic analysis. *Discourse Processes* 25, 259–284 (1998)
19. Névéal, A., Rogozan, A., Darmoni, S.: Automatic indexing of online health resources for a french quality controlled gateway. *Information Processing and Management* 42(3), 695–709 (2006)
20. Silberztein, M.: Nooj: an object-oriented approach. *Cahiers de la MSH Ledoux*, 359–369 (2004)

21. Diosan, L., Rogozan, A., Pecuchet, J.: Automatic alignment of medical vs. general terminologies. In: European Symposium on Artificial Neural Networks, Bruges, Belgium, pp. 487–492 (2008)
22. Vapnik, V.: *The Nature of Statistical Learning Theory*. Springer, Heidelberg (1995)
23. Loisel, A., Kotowicz, J.P., Chaignaud, N.: An issue-based approach to information search modelling. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds.) TSD 2008. LNCS (LNAI), vol. 5246, pp. 609–616. Springer, Heidelberg (2008)

An Intelligent Automatic Hoax Detection System*

Marin Vuković, Krešimir Pripuzić, and Hrvoje Belani

University of Zagreb, Faculty of Electrical Engineering and Computing,
Department of Telecommunications, Unska 3, HR-10000 Zagreb, Croatia
{marin.vukovic,kresimir.pripuzic,hrvoje.belani}@fer.hr

Abstract. Although they sometimes seem harmless, hoaxes represent not-negligible threat to individuals' awareness of real-life situations by deceiving them, and at the same time doing harm to the image of their organizations, which can lead to substantial financial losses. Spreading of hoaxes also influences the normal operating regime of networks and the efficiency of workers. In this paper we present an intelligent automatic hoax detection system based on neural networks and advanced text processing. In the developing of our system we use a database with real-life e-mail hoaxes, and an additional database with real-life e-mail messages. At the end we give brief experimental evaluation of the hoax detection system and comment the results.

Keywords: Hoax detection, e-mail classification, n-grams, self-organizing map, feed forward neural network, experimental evaluation.

1 Introduction

Hoaxes (the term origins from: hocus to trick) are more or less present throughout the entire history of mankind. Usual intention of hoax creator is to persuade or manipulate other people to do or prevent pre-established actions, mostly by using a threat or deception [1]. These intentions usually rely on empathy and abuse the human need for helping other people. Hoax creators want that their messages be read and forwarded to the largest possible number of victims. In today's world, hoaxing seems to find a fruitful ground in the new and emerging information and communication technologies (ICT), like e-mail, instant messaging, internet chats and mobile messaging.

Although hoaxes are not created to make technical damages to computer programs and operating systems, they can lead victims to damage their programs or systems, destroy reputation of them and their companies, coworkers and friends, or even to produce some financial losses. According to [2], which is considered to be the world's most widely quoted research on computer crime for years, various financial frauds result in an average reported loss of close to \$500,000 per company. The share of hoaxes in these frauds is not pointed out specifically, but the hoax-related losses, like insider abuse of networks and phishing, are mentioned and briefly analyzed in the

* This work was carried out within the research project "Content Delivery and Mobility of Users and Services in New Generation Networks", supported by the Ministry of Science, Education and Sports of the Republic of Croatia.

study. The spreading of hoaxes also influences the normal operating regime of network and efficiency of workers. Therefore, taking care of hoaxes is very important in order to minimize their impact on the different security-related frauds and misuses.

In science of memetics, a controversial research field that transcends psychology, biology, anthropology, and cognitive science, the term of hoaxes is sometimes referred to as a “virus of the mind” [3], mainly because of its ability to self-replicate, adapt, mutate, and persist in the human mind. They stay in the human mind and provoke the “infected” individuals to pass them to other individuals.

The next section surveys related work in the fields of text classification and hoax detection. The third section proposes a new hoax detection system based on artificial neural networks, while the fourth section describes a pre-processing of text, which is a very important part of the whole system. The fifth section presents and comments results of an experimental evaluation of implemented automatic hoax detection system. The sixth section gives conclusion and future work.

2 Related Work

Although the main definition of hoaxes origins from e-mail communication [4], in general it is every electronic message that contains bogus information with malicious intention to mislead its receiver. Such bogus information can be represented as textual, graphical, audio, and/or other multimedia content, e.g. fake virus warnings and various photo manipulations. In this work we focus on automatic detection of textual hoaxes, like e-mails [5], short message service (SMS) messages [6], and messages in Internet chats and forums.

When talking about unwanted e-mail messages, it is necessary to distinguish between hoaxes and unsolicited commercial e-mails, known as spam. Spam is, in its nature, created in order to sell a certain product or service, and therefore the usual expectation is that spam offers information which is exaggerate and not entirely true. On the other hand, hoax has a purpose of deceiving an average user and making him to believe in fake information it provides. A significant research work done in [7] addresses technological, organizational, behavioral, and legislative anti-spam measures, which unfortunately cannot be automatically applied to hoaxes.

The research area of automatic hoax detection gained a significant interest in the last decade, but with the partial results and solutions that are based on different approaches, e.g. heuristics, traffic analysis, etc. [1]. Authors of [4] developed a service that receives and evaluates e-mail messages that users forward when they suspect a hoax. Their approach and results are interesting, but not applicable for real-time hoax detection, in which we are interested. Furthermore, it is more difficult to detect if an unsuspected message is hoax.

For the classification of text, mostly unsupervised methods such as self-organizing maps (SOM) are used [8] [9], but there are also examples of using supervised learning methods [10] [11] [12]. The SOM architecture is very appropriate for this purpose, because it is able to classify the text according to the similarity of input patterns which represent the e-mail text. However, in order to avoid poor classification results, the SOM input patterns, which represent the text, must be coded in a proper manner. Since we are interested in distinguishing hoaxes from regular e-mails, using only

SOM, as done in [8] and [9], is not appropriate because it would classify a text into several clusters that contain both hoaxes and regular e-mails, depending on the input pattern coding scheme. Authors of [12] used only supervised learning for classification of e-mail messages to several groups. Since we are trying to distinguish hoaxes from regular e-mails only, our system can be more precise in performing its task. Besides the detection of hoax messages, we further improve our system to classify detected hoaxes as well. This is appropriate for various applications, such as improving the system in means of automatic learning and dealing with false positives which is discussed in the further text. Therefore, our solution combines both supervised and unsupervised learning. The supervised learning is used for distinguishing hoaxes from regular e-mails while unsupervised learning is used for hoax message classification.

3 Hoax Detection System

Our hoax detection system is composed of several modules as shown in Figure 1. After the preprocessing of e-mail content, which is described in the next section, the vector containing numeric representation of a single e-mail is presented to the hoax detector module.

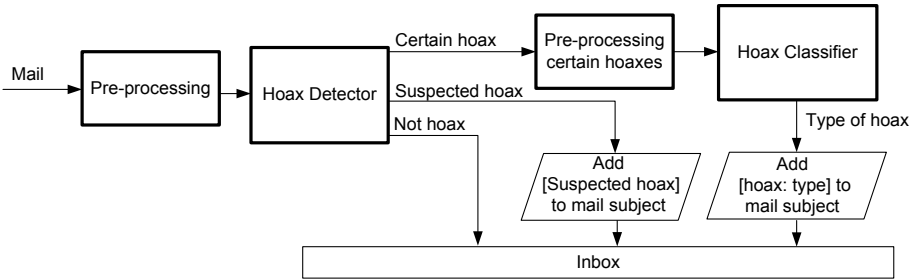


Fig. 1. Hoax Detection System

The hoax detector module is realized by a feed-forward artificial neural network whose task is to distinguish regular e-mails from e-mails containing hoaxes (or hoax e-mail in short). In order to accomplish this task, the network had to be learned with both regular and hoax e-mails using supervised learning technique. To train and test the neural networks we have used a collection of 298 hoax and 1370 regular e-mail messages. These messages were in the plain text form. The size of the hoax collection was 382 kilobytes, while the size of our mail collection was 2.92 megabytes. Learning set was composed as follows:

$$\text{Input: } (ntf_{1,d} \cdot idfb_{1,d}), (ntf_{2,d} \cdot idfb_{2,d}) \dots (ntf_{1369,d} \cdot idfb_{1369,d}),$$

$$\text{Output: } (hoax, \neg hoax),$$

where $idfb_{i,d}$ denotes the inverse document frequency in the both collections.

The learning was performed using backpropagation algorithm on the network consisting of the following three layers:

- Input layer with 1369 input neurons,
- Hidden layer with 100 neurons, and
- Output layer consisting of two neurons.

The input layer size corresponds to the total number of important n -grams, which are obtained in the pre-processing phase as described in the next section. The purpose of the two output neurons is to see whether an incoming e-mail is certainly hoax, is suspected to be hoax, or is certainly not hoax. This functionality is achieved by activating only one neuron in the case when network concludes that an e-mail is certainly hoax, and activating other neuron otherwise. However, if the network is unsure, both neurons will be activated with low certainty and further checking is required. In this context, we define low certainty as activation of less than 0.75. Furthermore, if only one neuron is activated with certainty lower than 0.5 we also conclude that further checking is required, in order to prevent false positive and/or false negative results.

The learning set consisted of 1668 patterns, of which 298 were known hoax e-mails, and others were regular e-mail messages. Once the learning is finished, the network is able to decide whether a pattern at the input is certainly hoax, is suspected to be hoax, or is certainly not hoax.

As we can see in Figure 1, if an incoming e-mail at the input of hoax detector module is recognized as certainly not hoax, it will be forwarded to the inbox. If an e-mail is suspected to be hoax, the phrase „suspected hoax” will be concatenated to its subject. In this way, the reader will be informed about the suspicion and will be able to proceed to the mail with caution. Additionally, if enabled, our system offers an option that users read and categorize e-mails as hoaxes or not. This way, any user can enhance the system performance by expanding the known hoax database. Finally, if the hoax detector is sure that the e-mail is hoax, the system is able to classify it in a group with similar hoaxes. This approach has the two main advantages:

1. The system continues to build the hoax database which is very useful for further system enhancements,
2. Users can see that a hoax belongs to a certain group of wide-spread, known hoaxes, thus improving general consciousness about hoaxes and risks regarding them.

The main task of hoax classifier is to classify hoax e-mails. In order to do so, all such e-mails must be pre-processed as described in the next section, in such a way that the input vector contains only n -grams from collection of hoax e-mails, and not from the collection of regular e-mail messages. A self organizing map is used as the hoax classifier. As this type of network architecture requires unsupervised learning, the patterns do not have to include an output of the SOM. The patterns were formed as follows:

$$(ntf_{1,d} \cdot idfh_{1,d}), (ntf_{2,d} \cdot idfh_{2,d}) \dots (ntf_{1880,d} \cdot idfh_{1880,d}),$$

where $idfh_{i,d}$ denotes the inverse document frequency in the hoax collection.

The SOM used for classification has 1880 input layer neurons and 20 neurons in the output layer. The training is done with Kohonen algorithm with the training set

consisting of 298 known hoax messages. The network results are interpreted by winner neuron (*Winner Takes All*). Our experimental results prove this to be suitable for the purpose, because the goal of the SOM is to classify a hoax in the group with the most similar hoaxes.

4 Text Pre-processing

As we explained in the previous section, we use two different neural networks in the system: a self-organizing map (SOM) and a feed-forward neural network. We use text processing methods to reduce the number of input neurons in both of these networks. This is a very important step for the whole system because it directly improves its performance by significantly reducing the number of input neurons in the neural networks. In this section we explain the text processing approach we use as a pre-processing step in our system.

Many of the regular e-mail and hoax messages in our collections were written in two or even more languages, but mostly in English or/and Croatian. Because of the mixture of different languages in a single document, we could not use stemming and lemmatization which are the standard text processing methods. In short, stemming usually refers to a crude heuristic process that chops off the ends of words in the hope of achieving reduction of a word to a common base form, while lemmatization achieves the same goal with the use of vocabulary [13] and morphological analysis of words [14]. Instead of these methods, we use n -gram (i.e. k -gram) tokenization [14], which is a process of breaking text to n -grams, where an n -gram is a sequence of n characters. For example, sentence “*Crni cvorak skakuce na širokoj grani*” breaks into the following 3-grams: {*crn, rni, cvo, vor, ora, rak, ska, kak, aku, kuc, uce, sir, iro, rok, oko, koj, gra, ran, ani*}.

Actually, we have to normalize text before n -gram tokenization. This is done by removing capitalization, punctuation and diacritics. Removing of capitalization is the first necessary step in our text normalization. This way we ignore the difference among same words that are capitalized differently. For example, we treat words “*Crni*” and “*crni*” equally. The second step in the normalization is to remove special characters and punctuation from text. It is very common to write text in Croatian without using the diacritics. For example, writing “*Crni cvorak skakuce na širokoj grani.*” instead of the proper sentence „*Crni čvorak skakuće na širokoj grani.*” Therefore, the removing of diacritics is the third necessary step in the normalization. We summarize our text processing approach as follows:

1. Text Normalization
 - a. Removing of capitalization
 - b. Removing of special characters and punctuation
 - c. Removing of diacritics
2. Text Tokenization
 - a. n -gram tokenization

In the process of tokenization, we limit the size of n -grams to $n = 3 \dots 8$. The shorter n -grams are too general and we have excluded them because their entropy is very low.

On the other side, the longer n -grams are too specific for belonging document, and we have excluded them because they are not appropriate for the generalization among different documents. The total number of such n -grams in the both collections was 79448.

This is quite a large number of n -grams to directly use them as inputs of our neural networks. That is why we need to reduce their number such that we include only those n -grams that are actually important for pre-processing of text. For this purpose we use document frequency df_t , defined to be the number of documents in a collection that contains term t . As we can see in Figure 1, we use two different pre-processing modules: mail pre-processing and pre-processing of certain hoaxes. The only difference among these two modules is in a different selection of the **important n -grams**.

In the mail pre-processing, we discard n -grams that are rare in the hoax collection $df_t(hoax) < 15$ or very common in the both collections $df_t(hoax+mail) < 50$. The number of important n -grams (and input nodes in the first neural networks) obtained in this way is 1369. We drop common n -grams (stop words) because they cannot help us in distinction among different documents, while rare n -grams (discriminative words) cannot help the neural network in generalization. Because of the same reason, in the pre-processing of certain hoaxes, we discard n -grams that are rare $df_t(hoax) < 10$ or very common $df_t(hoax) < 20$ in the hoax collection. This time the number of obtained important n -grams is 1880. It is important to notice that the thresholds are lower in the second pre-processor. The main reason for this is the second neural network, which is a classifier, and therefore we need more rare n -grams as inputs to distinguish between different hoaxes.

At this point, we will explain how we create input values for the neural networks in the system. For an incoming e-mail, we apply the first two steps of our text processing approach (i.e. the normalization and tokenization). Then for each important n -gram in the e-mail we calculate product of its *normalized term frequency* and *inverse document frequency* values:

$$ntf_{t,d} \cdot idf_{t,d} = \left(a + (1 - a) \frac{tf_{t,d}}{mtf_d} \right) \cdot \log \frac{N}{df_t},$$

where $a = 0.4$ is the standard value of smoothing term, N is a number of documents in the related collection, $tf_{t,d}$ is n -gram frequency in the e-mail, and $mtf_d = \max_{t \in d} tf_{t,d}$ is maximal n -gram frequency of all n -grams in the e-mail. These $ntf_{t,d} \cdot idf_{t,d}$ values are inputs for the neural networks.

5 Experimental Evaluation

First we evaluate the performance of the hoax detector module. After training the module is capable to distinguish all hoaxes from regular e-mails, which were contained in the training set, as expected. However, if a new e-mail is received at an input, it is classified according to its similarity with the “known” e-mails. Thus, the results are shown in the Table 1. Obviously, the main issue concerns suspected hoaxes, i.e. e-mails which may or may not be hoaxes. This is because they contain words which are common in both hoaxes and regular e-mails.

Table 1. Performance results of the hoax detector module

False positives	Suspected hoax	False negatives	Correct
4,90%	19,70%	1,54%	73,86%

The evaluation results of the hoax classifier module are the following. After training the SOM, the known hoax messages are divided in 20 groups, analogue to number of output neurons. The results for the four most common hoax groups are presented in Table 2. The outcome of the classifier could be altered by modifying the number of SOM output neurons if necessary. However, this has proven to be suitable for our purpose.

Table 2. The four most common hoax groups

Group ID	Hoax theme	number of messages	percentage in hoax dataset
Group 6	Chained letters – prayers (in Croatian)	26	8,70%
Group 9	Chained letters – prayers (in English)	21	7%
Group 11	Asking help for surgery (in Croatian)	20	6,70%
Group 17	Warning recipients about something (in Croatian)	20	6,70%

6 Conclusion and Future Work

This paper proposes intelligent hoax detection system based on artificial neural networks for which the data have to be pre-processed using information retrieval methods. In experimental evaluation, we showed that it is possible to detect and classify hoax messages successfully, at least to some extent. The proposed system has the ability to distinguish and classify hoax messages by comparing them against known hoax messages which usually contain similar patterns. However, if a new hoax message would appear, which does not have any similarity with the ones contained in the training set, the proposed system would not be able to detect it. The main issue with hoax messages in general is that they could be very similar to regular e-mail messages and it is difficult to distinguish whether their content is true or not, even to a human. As future work, we plan to develop a technique which could further evaluate the message content thus lowering the number of false positives and, especially, suspected hoax messages.

Furthermore, implemented system can also be applied to automatic hoax detection in SMS messages, which becomes one of raising issues in the evolving field of value added services (VAS) in telecommunications. Nevertheless, this kind of system evaluation would require an easy-manageable dataset of SMS hoaxes, as well as regular SMS messages, which is not trivial to acquire.

References

1. Hernandez, J.C., Hernandez, C.J., Sierra, J.M., Ribagorda, A.: A First Step towards Automatic Hoax Detection. In: Proceedings of the International 36th Annual Carnahan Conference on Security Technology, pp. 102–114. IEEE, Piscataway (2002)
2. Richardson, R.: CSI Computer Crime & Security Survey. Computer Security Institute, San Francisco, CA (2008), <http://www.gocsi.com/>
3. Brodie, R.: Virus of the Mind: The New Science of the Meme, 1st edn. Integral Press, USA (1995)
4. Petković, T., Kostanjčar, Z., Pale, P.: E-Mail System for Automatic Hoax-Recognition. In: XXVII. International Convention MIPRO 2005 Bd. CTS & CIS, Opatija, Croatia, pp. 117–121 (2005) ISBN 953–233–012–7
5. Sakkis, G.: Learning How to Tell Ham from Spam. *Crossroads* 11(2) (2004)
6. SMS Hoax Causes Traffic Congestion. *textually.org: all about texting, SMS and MMS* (2009), <http://www.textually.org/textually/archives/2005/08/009494.htm> (accessed on: March 2009)
7. Schryen, G.: *Anti-Spam Measures - Analysis and Design*. Springer, Heidelberg (2007)
8. Kim, H.-D., Cho, S.-B.: Application of Self-Organizing Maps to Classification and Browsing of FAQ E-mails. In: Kowalczyk, R., Loke, S.W., Reed, N.E., Graham, G. (eds.) *PRICAI-WS 2000. LNCS (LNAI)*, vol. 2112, pp. 44–55. Springer, Heidelberg (2001)
9. Merkl, D., Rauber, A.: Document Classification with Unsupervised Artificial Neural Networks. In: Crestani, F., Pasi, G. (eds.) *Soft computing in information retrieval*, pp. 102–121. Physica-Verlag, Heidelberg (2000)
10. Jevtic, D., Car, Ž., Vukovic, M.: Location Name Extraction for User Created Digital Content Services. In: Apolloni, B., Howlett, R.J., Jain, L. (eds.) *KES 2007, Part I. LNCS (LNAI)*, vol. 4692, pp. 623–630. Springer, Heidelberg (2007)
11. Cui, B., Mondal, A., Shen, J., Cong, G., Tan, K.-L.: On Effective E-mail Classification via Neural Networks. In: Andersen, K.V., Debenham, J., Wagner, R. (eds.) *DEXA 2005. LNCS*, vol. 3588, pp. 85–94. Springer, Heidelberg (2005)
12. Clark, J., Koprinska, I., Poon, J.: A Neural Network Based Approach to Automated E-mail Classification. In: Proceedings of the IEEE/WIC International Conference on Web Intelligence (WI 2003), Halifax, Canada, pp. 702–705. Computer Society Press (2003)
13. Pripužić, K., Huljenić, D., Carić, A.: Vocabulary Development for Event Notification Services. In: Proceeding of The International Conference on Software, Telecommunications and Computer Networks SoftCOM 2004, Split-Dubrovnik-Venice, Croatia-Italy (2004)
14. Manning, C.D., Raghavan, P., Schütze, H.: *Introduction to Information Retrieval*. Cambridge University Press, New York (2008)

Web User Session Reconstruction with Back Button Browsing

Robert F. Dell¹, Pablo E. Román^{2,*}, and Juan D. Velásquez²

¹ Naval Postgraduate School, Operations Research Department,
Monterey, California, USA

² University of Chile, Department of Industrial Engineering,
República 701, Santiago, Chile

Abstract. A web user session, the sequence of pages a user visits at a web site, is valuable data used in many e-business applications but privacy concerns often limit their direct retrieval. A web server log file provides an approximate way of constructing user sessions without privacy concerns. It is only approximate because the same IP address as recorded in the web log often contains the requests of several concurrent users without each user being uniquely identified. Additionally, a user's activation of the back and forward browser button is often not recorded in the web log because, in most cases, the browser retrieves the page from its own cache. We present an integer program to construct user sessions (sessionization) from web log data that includes the possible use of the back button. We present sessionization results on web log data from an academic web site and compare sessions constructed with and without the option of sessions with the back button.

1 Introduction

A web server log file is a list of registers that collectively provide valuable (but incomplete) data on user activities at a web site. What the web log fails to directly capture is an individual user's session, the sequence of pages a user visits at a web site, that are used as a primary input to web mining [1]. Each web log register typically includes the access time of an object, its URL, the IP address of the user, the referrer corresponding to last visited URL and the agent field identifying the user's browser. Unfortunately, the same IP address as recorded in the web log often contains the requests of several concurrent users without each user being uniquely identified because the network address translation (NAT) used by most internet service providers share the same IP number for different clients. Additionally, a user's activation of the back and forward browser button is often not recorded in the web log because, in most cases, the browser retrieves the page from its own cache.

Web browsing studies have identified up to 30% of the pages visited during a session are obtained using the back button, while fewer than (2%) use the

* Corresponding author.

forward button [2]. Prior work on sessionization (constructing sessions) has not considered the use of the back button; most have relied on simple heuristics [3,4,5]. These heuristics group sessions by the same IP and agent so that each session's duration does not exceed more than 30 minute. Others [6] additionally use the referrer field but we have found this field is often not available due principally to client side blocking. Time driven heuristics have been popular for web mining applications [1], but we have achieved better results using optimization models [7]. In this paper, we extend our prior work to consider the use of the back button.

The rest of this paper is organized as follows. Section 2 presents our integer program. Section 3 presents our test data and results. Section 4 provides some conclusions and suggests future research.

2 Integer Program for Sessionization Considering the Back Button

Sessions are reconstructed as ordered list of log registers. For our notation, each index r identifies a unique register, each index s identifies a unique user session, and the index o is the ordered request of a register during a session. We consider the possible use of the back button once for each register because several continued back events are rare in web browsing [2] and the forward button is scarcely used [8]. We use the binary variable Y_{ros} that has value one if log register r is assigned as the oth request during session s and repeats in the $o + 2$ position by using the back button from the $o + 1$ position. The use of variable Y_{ros} is the primary difference from the formulation found in our prior work ([7]). The binary variable X_{ros} has value one if log register r is assigned as the oth request during session s without being repeated using the back button. We present the integer programming formulation below in NPS standard format [9] (a define before use format for writing optimization models).

2.1 Indices

o	Order of a log register visit during a session. The cardinality defines the maximum length of a session.
p, p'	Web page.
r, r'	Web log register.
s	Web user session.

2.2 Index Sets

$r' \in bpage_r$	The set of registers that can be the register immediately before register r in the same session. Based on: <ul style="list-style-type: none"> - the referrer URL for register r (if available) - pages available from the page of register r in one click - pages available from the page of register r in one click
------------------	--

- IP address matching of register r and register r'
- agent matching of register r and register r'
- time of register r and register r' .

Of course, r can not occur before r' but we assume a user defined minimum and maximum time between two consecutive registers in the same session.

$r \in first$ set of registers that must be first in a session.

2.3 Data [units]

Used to produce the index sets above:

- $time_r$ the time of register r [seconds].
- ip_r the IP address for register r .
- $agent_r$ the agent for register r .
- $page_r$ the page for register r .
- $\overline{mtp}, \underline{mtp}$ the min, max time between pages in a session [seconds].
- $\overline{adjacent}_{p,p'}$ one if a page p' can be reached in one click from page p .

Used in formulation:

- C_{ros} the objective function weight of having register r assigned to the oth position in session s .

2.4 Binary Variables

- X_{ros} 1 if log register r is assigned as the oth request during session s and zero otherwise.
- Y_{ros} 1 if log register r is assigned as the oth and $(o + 2)th$ request during session s indicating that a back button action was performed from the $(o + 1)th$ position, and zero otherwise.

2.5 Formulation

Maximize $Z = \sum_{ros} \{C_{ros}X_{ros} + (C_{ros} + C_{r,o+2,s})Y_{ros}\}$

Subject to:

$$\sum_{os} X_{ros} + Y_{ros} \leq 1 \quad \forall r \quad (1)$$

$$\sum_o \{X_{ros} + Y_{ros} + Y_{r,o-2,s}\} \leq 1 \quad \forall o, s \quad (2)$$

$$X_{r,o+1,s} + Y_{r,o+1,s} \leq \sum_{r' \in bpage_r} \{X_{r',o,s} + Y_{r',o,s} + Y_{r',o-2,s}\} \quad \forall r, o, s \quad (3)$$

$$Y_{r,o,s} \leq \sum_{r' | r \in bpage_{r'}} \{X_{r',o+1,s} + Y_{r',o+1,s} + Y_{r',o-1,s}\} \quad \forall r, o, s \quad (4)$$

$$X_{ros} \in \{0, 1\}, \quad Y_{ros} \in \{0, 1\} \quad \forall r, o, s,$$

$$X_{ros} = 0, \forall r \in first, o > 1, s$$

The objective function expresses the total reward for sessions where it uses the same coefficients from earlier studies [7] extended for the possibility of back button usage (Y_{ros}). The term $\sum_{ros} (C_{ros} + C_{r,o+2,s})Y_{ros}$ includes coefficients for both the o and $o + 2$ ordered request. Constraint set (1) ensures each register is used at most once. Constraint set (2) restricts each session to have at most one register assigned for each ordered request. Constraint set (3) ensures the proper ordering of registers in the same session. Constraint set (4) ensures a register follows in the $o + 1$ position when $Y_{ros} = 1$. To improve solution time, we can fix (or eliminate) a subset of these binary variables to zero ($Y_{ros} = 0$ and $X_{ros} = 0, \forall r \in first, o > 1, s$). After forming the set $bpage_r$, the set $first$ is easily found ($r \in first$ if $bpage_r = \emptyset$).

3 Web Log Processing

We use the same web log used in [7]. It consists of 3,756,006 raw register collected from the Industrial Engineering Department of the University of Chile (<http://www.dii.uchile.cl>) web site during April 2008. After filtering, we obtain a total of 102,303 clean registers of static html pages as well some dynamic pages (php and jsp) corresponding to 172 different pages with 1,228 links between them. Of these, 9,044 registers correspond to visits to the root page of the site. The available web log fields include the IP address, time, URL, and agent. The referrer field is not available. 98 percent of the IP addresses have less than 50 register for the entire month. 84 percent of the IP addresses visit three or less different pages for the entire month. Information is stored in a relational database ensuring data consistency.

We use a subset of the clean registers, selected by IP address and web page diversity, from [7]. The measure of diversity is entropy, $S = \sum_p f_p \text{Log}_N(1/f_p)$, where f_p is the frequency of page p occurrence over all register entries for the same IP address and N is the number of unique pages visited by the same IP address. S takes values from zero to one. When the value of S is near zero, most register entries are for the same page, if the value is near one all the pages are visited with similar frequency. IP addresses with high diversity and a high number of registers are selected because we have found these are the most interesting (and most difficult to solve) for sessionization. Selecting the IP addresses with more than 50 registers and S greater than 0.5 results in 130 IP addresses with 17,709 registers (17.3% of the total number of registers). The log file is split into 403 chunks when partitioning these registers by IP and agent such that each chunk has at least 50 registers and $\overline{mtp} = 300$.

All computation is done using 1.6Ghz dual-core PC with two Gbs of RAM. We generate the integer program using GAMS and solve it using CPLEX version 10.1.0 [10], controlled by a script program.

3.1 The Sessionization Processing

We use objective function coefficients $C_{ros} = 3/2 \text{Log}(o) + (o - 3)^2/12o$, based on prior work with these values [7]. The motivation for this expression was

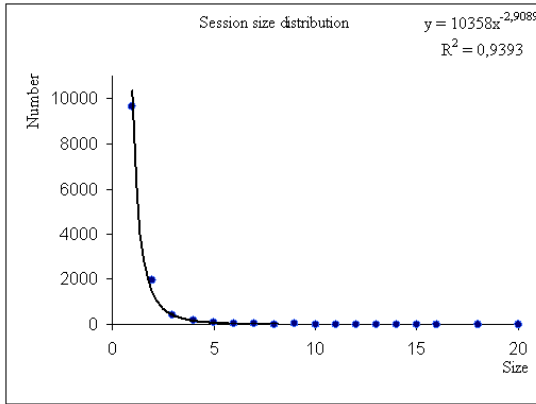


Fig. 1. Session size distribution

$\text{Log}(1/P_o)$, where P_o is the inverse Gaussian probability reported in [11] for sessions of size o . Figure 1 shows how many of each session size is found by our integer program.

It requires 7hr 15min to solve the 403 chunks (with $mtp = 0$ and $\overline{mtp} = 300$, and a maximum session size of 20). Where for each integer program, we set the maximum time limit to 300 seconds and the *relative gap* to one percent. With such limits, the CPLEX solver terminates when it has a solution guaranteed to be with one percent of optimal or it reaches a 300 second (five minute) limit. If it reaches a 300 second limit, it provides the best solution it has found by that time. The different integer programs vary from 7601 to 704, 801 variables and from 402 to 572, 490 constraints. Over 86% (345 out of 403) of the chunks obtained a solution within one percent of optimal. The average relative gap of these 345 chunks was 0.15%. The average generation and solution time for these chunks was 8.7 seconds. The 300 second limit was reached in 58 out of the 403 chunks.

We do not know how well the sessions match reality because we do not know the actual sessions that give rise to a specific web server log. One measure of quality previously used in [7] is how well the distribution of sessions matches to the empirical observed power law distribution (approximation to the inverse Gaussian) [11,12], using linear regression on log scale of number of session versus the session size. The results with the possibility of a hidden step (back button) on sessions give a correlation value of $R^2 = 0.94$ and a standard error of $err = 0.64$ for 12, 491 sessions as shown in Figure 1. This is slightly better than results using a time-oriented heuristics [5] (without the use of a back button), a correlation value of $R^2 = 0.91$ and a standard error of 0.64.

3.2 Comparing to Sessions without Using the Back Button

Earlier work on sessionization does not consider the possible use of the back button. Figure 2 provides a comparison of the session size using the model of

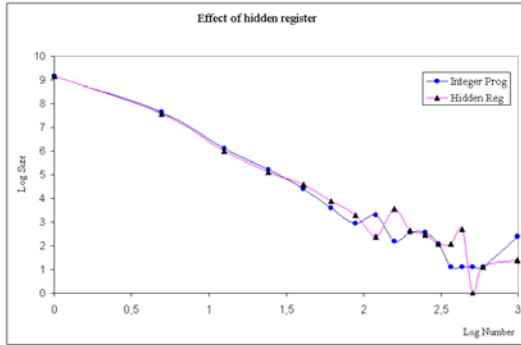


Fig. 2. Comparing the number of sessions (in log scale) with and without the use of the back button

section 2.5 with our earlier sessionization model [7]. We find the number of sessions of up to $size = 7$ are similar but the use of the back button produces 17% more sessions of size greater than 7.

There are a total of 12,491 sessions found when allowing the use of the back button. Of these, 860 sessions had at least one register that was repeated by using the back button. Of the 2837 sessions of size greater than 2, 32% of the sessions had at least one register that was repeated by using the back button. This compares favorably with the 30% use of the back button found in previous studies [2].

4 Conclusions and Future Research

We present a new approach for sessionization using an integer program that includes the possible use of a browser back button. We compare sessionization results with and without the possible use of the back button and find similar results for shorter sessions. We find more longer sessions when allowing the possibility of the browser back button and find these session have back button use similar to what has been empirically observed.

Future work includes extensions for sites with frames that generalize the concept of a web page to a frameset [13].

Acknowledgement

This work has been partially supported by the National Doctoral Grant from Conicyt Chile and by the Chilean Millennium Institute of Complex Engineering Systems.

References

1. Velásquez, J.D., Palade, V.: Adaptive web sites: A knowledge extraction from web data approach. IOS Press, Amsterdam (2008)
2. Tauscher, L., Greenberg, S.: Revisitation patterns in world wide web navigation. In: Proc. of the Conference on Human Factors in Computing Systems, Atlanta, USA, March 1997, pp. 22–27 (1997)
3. Berendt, B., Hotho, A., Stumme, G.: Data preparation for mining world wide web browsing patterns. *Journal of Knowledge and Information Systems* 1(1), 5–32 (1999)
4. Cooley, R., Mobasher, B., Srivastava, J.: Towards semantic web mining. In: Horrocks, I., Hendler, J. (eds.) ISWC 2002. LNCS, vol. 2342, pp. 264–763. Springer, Heidelberg (2002)
5. Spiliopoulou, M., Mobasher, B., Berendt, B., Nakagawa, M.: A framework for the evaluation of session reconstruction heuristics in web-usage analysis. *Inform Journal on Computing* 15(2), 171–190 (2003)
6. Srivastava, J., Cooley, R., Deshpande, M., Tan, P.: Web usage mining: Discovery and applications of usage patterns from web data. *SIGKDD Explorations* 2(1), 12–23 (2000)
7. Dell, R., Róman, P., Velásquez, J.: Web user session reconstruction using integer programming. In: Proc. of The 2008 IEEE/WIC/ACM International Conference on Web Intelligence, Sydney, Australia, December 2008, pp. 385–388 (2008)
8. Catledge, L.D., Pitkow, J.E.: Characterizing browsing strategies in the world-wide web. In: *Computer Networks and ISDN Systems*, pp. 1065–1073 (1995)
9. Brown, G.G., Dell, R.F.: Formulating integer linear programs: A rogues’ gallery. *Inform Transactions on Education* 7(2), 1–13 (2007)
10. GAMS Development Corporation: Gams/cplex, Solver CPLEX for GAMS (2008), <http://www.gams.com/dd/docs/solvers/cplex.pdf>
11. Huberman, B., Pirolli, P., Pitkow, J., Lukose, R.M.: Strong regularities in world wide web surfing. *Science* 280(5360), 95–97 (1998)
12. Vazquez, A., Oliveira, J.G., Dezso, Z., Goh, K.I., Kondor, I., Barabasi, A.L.: Modeling bursts and heavy tails in human dynamics. *Physical Review E* 73(3), 036127 (2006)
13. Mobasher, B., Berent, B., Spiliopoulou, M., Wiltshire, J.: Measuring the accuracy of sessionizers for web usage analysis. In: *Proceedings of the Web Mining Workshop at the First SIAM ICDM* (2001)

Fast Time Delay Neural Networks for Detecting DNA Coding Regions

Hazem M. El-Bakry¹ and Mohamed Hamada²

¹ Faculty of Computer Science & Information Systems,
Mansoura University, Egypt
helbakry20@yahoo.com

² University of Aizu
Aizu Wakamatsu, Japan
Hamada@u-aizu.ac.jp

Abstract. In this paper, a new approach for fast information detection in DNA sequence has been presented. Our approach uses fast time delay neural networks (FTDNN). The operation of these networks relies on performing cross correlation in the frequency domain between the input data and the input weights of neural networks. It is proved mathematically and practically that the number of computation steps required for the presented FTDNNs is less than that needed by conventional time delay neural networks (CTDNNs). Simulation results using MATLAB confirm the theoretical computations.

1 Introduction

It is important to detect some specific information in DNA sequence such as protein coding regions [13-16]. Recently, time delay neural networks have shown very good results in different areas such as automatic control, speech recognition, blind equalization of time-varying channel and other communication applications. The main objective of this research is to reduce the response time of time delay neural networks. The purpose is to perform the testing process in the frequency domain instead of the time domain. Our approach was successfully applied for sub-image detection using fast neural networks (FNNs) as proposed in [1,2]. Furthermore, it was used for fast face detection [7,9], and fast iris detection [6]. Another idea to further increase the speed of FNNs through image decomposition was suggested in [5].

FNNs for detecting a certain code in one dimensional serial stream of sequential data were described in [4,5]. Compared with conventional neural networks, FNNs based on cross correlation between the tested data and the input weights of neural networks in the frequency domain showed a significant reduction in the number of computation steps required for certain data detection [1,2,3,5,7,8]. Here, we make use of our theory on FNNs implemented in the frequency domain to increase the speed of time delay neural networks. The idea of moving the testing process from the time domain to the frequency domain is applied to time delay neural networks. Theoretical and practical results show that the proposed FTDNNs are faster than CTDNNs. In section 2, our theory on FNNs for detecting certain data in one dimensional matrix is described. Experimental results for FTDNNs are presented in section 3.

2 Theory of FTDNNs Based on Cross Correlation in the Frequency Domain

Finding a certain pattern of information, in the incoming serial data, is a searching problem. First neural networks are trained to classify the required pattern from other examples and this is done in time domain. In pattern detection phase, each position in the incoming matrix is tested for presence or absence of the required information in DNA sequence. At each position in the input one dimensional matrix, each sub-matrix is multiplied by a window of weights, which has the same size as the sub-matrix. The outputs of neurons in the hidden layer are multiplied by the weights of the output layer. When the final output is high, this means that the sub-matrix under test contains the required information in DNA sequence and vice versa. Thus, we may conclude that this searching problem is a cross correlation between the incoming serial data and the weights of neurons in the hidden layer.

The convolution theorem in mathematical analysis says that a convolution of f with h is identical to the result of the following steps: let F and H be the results of the Fourier Transformation of f and h in the frequency domain. Multiply F and H^* in the frequency domain point by point and then transform this product into the spatial domain via the inverse Fourier Transform. As a result, these cross correlations can be represented by a product in the frequency domain. Thus, by using cross correlation in the frequency domain, speed up in an order of magnitude can be achieved during the detection process [1,2,3,4,5,7,8,9,14]. Assume that the size of the attack code is $1 \times n$. In attack detection phase, a sub matrix I of size $1 \times n$ (sliding window) is extracted from the tested matrix, which has a size of $1 \times N$. Such sub matrix, which may be an attack code, is fed to the neural network. Let W_i be the matrix of weights between the input sub-matrix and the hidden layer. This vector has a size of $1 \times n$ and can be represented as $1 \times n$ matrix. The output of hidden neurons $h(i)$ can be calculated as follows:

$$h_i = g \left(\sum_{k=1}^n W_i(k)I(k) + b_i \right) \tag{1}$$

where g is the activation function and $b(i)$ is the bias of each hidden neuron (i). Equation 1 represents the output of each hidden neuron for a particular sub-matrix I . It can be obtained to the whole input matrix Z as follows:

$$h_i(u) = g \left(\sum_{k=-n/2}^{n/2} W_i(k) Z(u+k) + b_i \right) \tag{2}$$

Eq.2 represents a cross correlation operation. Given any two functions f and d , their cross correlation can be obtained by:

$$d(x) \otimes f(x) = \left(\sum_{n=-\infty}^{\infty} f(x+n)d(n) \right) \tag{3}$$

Therefore, Eq. 2 may be written as follows [1]:

$$h_i = g(W_i \otimes Z + b_i) \tag{4}$$

where h_i is the output of the hidden neuron (i) and $h_i(u)$ is the activity of the hidden unit (i) when the sliding window is located at position (u) and $(u) \in [N-n+1]$.

Now, the above cross correlation can be expressed in terms of one dimensional Fast Fourier Transform as follows [1]:

$$W_i \otimes Z = F^{-1}(F(Z) \bullet F^*(W_i)) \quad (5)$$

Hence, by evaluating this cross correlation, a speed up ratio can be obtained comparable to conventional neural networks. Also, the final output of the neural network can be evaluated as follows:

$$O(u) = g\left(\sum_{i=1}^q W_o(i) h_i(u) + b_o\right) \quad (6)$$

where q is the number of neurons in the hidden layer. $O(u)$ is the output of the neural network when the sliding window located at the position (u) in the input matrix Z . W_o is the weight matrix between hidden and output layer.

The complexity of cross correlation in the frequency domain can be analyzed as follows:

1. For a tested matrix of $1 \times N$ elements, the 1D-FFT requires a number equal to $N \log_2 N$ of complex computation steps [13]. Also, the same number of complex computation steps is required for computing the 1D-FFT of the weight matrix at each neuron in the hidden layer.
2. At each neuron in the hidden layer, the inverse 1D-FFT is computed. Therefore, q backward and $(1+q)$ forward transforms have to be computed. Therefore, for a given matrix under test, the total number of operations required to compute the 1D-FFT is $(2q+1)N \log_2 N$.
3. The number of computation steps required by FTDNNs is complex and must be converted into a real version. It is known that, the one dimensional Fast Fourier Transform requires $(N/2) \log_2 N$ complex multiplications and $N \log_2 N$ complex additions [11]. Every complex multiplication is realized by six real floating point operations and every complex addition is implemented by two real floating point operations. Therefore, the total number of computation steps required to obtain the 1D-FFT of a $1 \times N$ matrix is:

$$\rho = 6((N/2) \log_2 N) + 2(N \log_2 N) \quad (7)$$

which may be simplified to:

$$\rho = 5N \log_2 N \quad (8)$$

4. Both the input and the weight matrices should be dot multiplied in the frequency domain. Thus, a number of complex computation steps equal to qN should be considered. This means $6qN$ real operations will be added to the number of computation steps required by FTDNNs.
5. In order to perform cross correlation in the frequency domain, the weight matrix must be extended to have the same size as the input matrix. So, a number of zeros

= (N-n) must be added to the weight matrix. This requires a total real number of computation steps = q(N-n) for all neurons. Moreover, after computing the FFT for the weight matrix, the conjugate of this matrix must be obtained. As a result, a real number of computation steps = qN should be added in order to obtain the conjugate of the weight matrix for all neurons. Also, a number of real computation steps equal to N is required to create butterflies complex numbers ($e^{-jk(2\pi n/N)}$), where $0 < K < L$. These (N/2) complex numbers are multiplied by the elements of the input matrix or by previous complex numbers during the computation of FFT. To create a complex number requires two real floating point operations. Thus, the total number of computation steps required for FTDNNs becomes:

$$\sigma = (2q+1)(5N\log_2N) + 6qN + q(N-n) + qN + N \tag{9}$$

which can be reformulated as:

$$\sigma = (2q+1)(5N\log_2N) + q(8N-n) + N \tag{10}$$

6. Using sliding window of size 1xn for the same matrix of 1xN pixels, q(2n-1)(N-n+1) computation steps are required when using CTDNNs for certain attack detection or processing (n) input data. The theoretical speed up factor η can be evaluated as follows:

$$\eta = \frac{q(2n-1)(N-n+1)}{(2q+1)(5N\log_2N) + q(8N-n) + N} \tag{11}$$

CTDNNs and FTDNNs are shown in Figures 1 and 2 respectively.

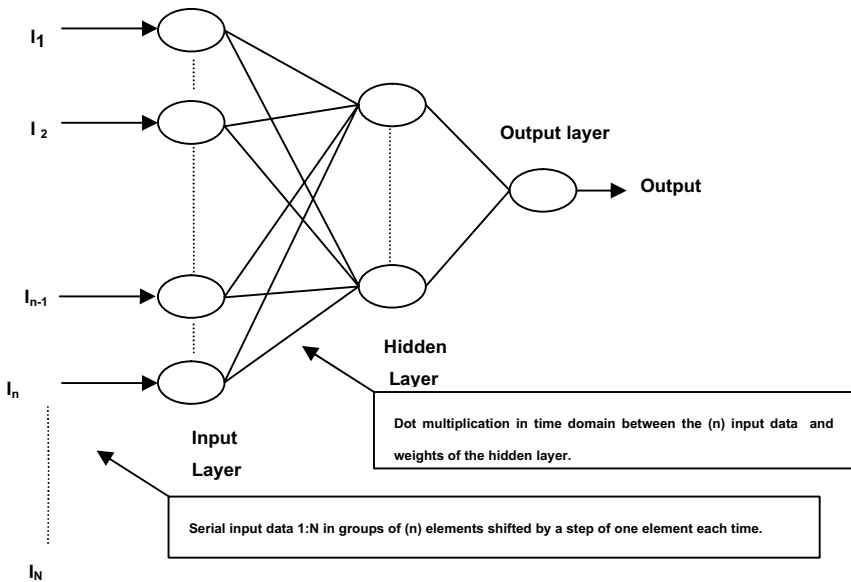


Fig. 1. Classical time delay neural networks

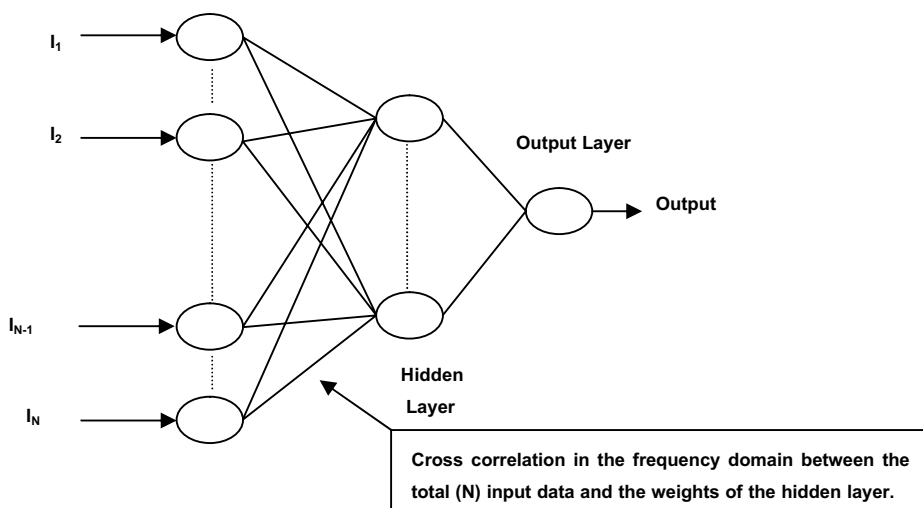


Fig. 2. Fast time delay neural networks

3 Fast Detection of Important Information in DNA Sequence by Using FTDNN

Time delay neural networks accept serial input data with fixed size (n). Therefore, the number of input neurons equals to (n). Instead of treating (n) inputs, the proposed new approach is to collect all the incoming data together in a long vector (for example $100 \times n$). Then the input data is tested by time delay neural networks as a single pattern with length L ($L=100 \times n$). Such a test is performed in the frequency domain as described in section II. The combined attack in the incoming data may have real or complex values in a form of one or two dimensional array. Complex-valued neural networks have many applications in fields dealing with complex numbers such as telecommunications, speech recognition and image processing with the Fourier Transform [3,8]. Complex-valued neural networks mean that the inputs, weights, thresholds and the activation function have complex values. In this section, formulas for the speed up ratio with different types of inputs (real /complex) will be presented. Also, the speed up ratio in case of a one and two dimensional incoming input matrix will be concluded. The operation of FTDNNs depends on computing the Fast Fourier Transform for both the input and weight matrices and obtaining the resulting two matrices. After performing dot multiplication for the resulting two matrices in the frequency domain, the Inverse Fast Fourier Transform is determined for the final matrix. Here, there is an excellent advantage with FTDNNs that should be mentioned. The Fast Fourier Transform is already dealing with complex numbers, so there is no change in the number of computation steps required for FTDNNs. Therefore, the speed up ratio in case of complex-valued time delay neural networks can be evaluated as follows:

1) In case of real inputs

A) For a one dimensional input matrix

Multiplication of (n) complex-valued weights by (n) real inputs requires (2n) real operations. This produces (n) real numbers and (n) imaginary numbers. The addition of these numbers requires (2n-2) real operations. The multiplication and addition operations are repeated (N-n+1) for all possible sub matrices in the incoming input matrix. In addition, all of these procedures are repeated at each neuron in the hidden layer. Therefore, the number of computation steps required by conventional neural networks can be calculated as:

$$\theta=2q(2n-1)(N-n+1) \tag{12}$$

The speed up ratio in this case can be computed as follows:

$$\eta = \frac{2q(2n-1)(N-n+1)}{(2q+1)(5N\log_2N)+q(8N-n)+N} \tag{13}$$

Practical speed up ratio for searching short successive (n) data in a long input vector (L) using complex-valued time delay neural networks is shown in Figure 3. This has been performed by using a 700 MHz processor and MATLAB.

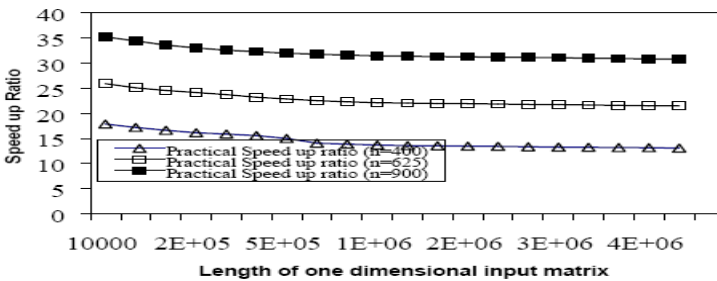


Fig. 3. Practical speed up ratio when using FTDNNs in case of one dimensional real-valued input matrix and complex-valued weights

B) For a two dimensional input matrix

Multiplication of (n²) complex-valued weights by (n²) real inputs requires (2n²) real operations. This produces (n²) real numbers and (n²) imaginary numbers. The addition of these numbers requires (2n²-2) real operations. The multiplication and addition operations are repeated (N-n+1)² for all possible sub matrices in the incoming input matrix. In addition, all of these procedures are repeated at each neuron in the hidden layer. Therefore, the number of computation steps required by conventional neural networks can be calculated as:

$$\theta=2q(2n^2-1)(N-n+1)^2 \tag{14}$$

The speed up ratio in this case can be computed as follows:

$$\eta = \frac{2q(2n^2 - 1)(N - n + 1)^2}{(2q + 1)(5N^2 \log_2 N^2) + q(8N^2 - n^2) + N} \quad (15)$$

Practical speed up ratio for detecting $(n \times n)$ real valued submatrix in a large real valued matrix $(N \times N)$ using complex-valued time delay neural networks is shown in Fig. 4. This has been performed by using a 700 MHz processor and MATLAB.

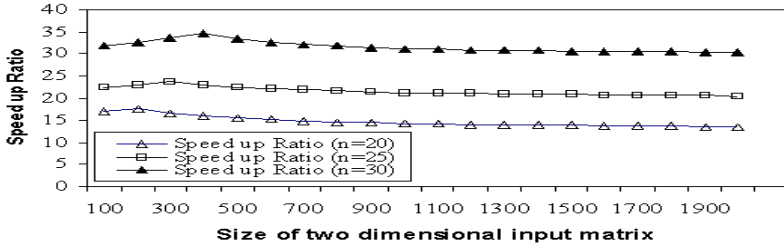


Fig. 4. Practical speed up ratio when using FTDNNs in case of two dimensional real-valued input matrix and complex-valued weights

2) In case of complex inputs

A) For a one dimensional input matrix

Multiplication of (n) complex-valued weights by (n) complex inputs requires $(6n)$ real operations. This produces (n) real numbers and (n) imaginary numbers. The addition of these numbers requires $(2n-2)$ real operations. Therefore, the number of computation steps required by conventional neural networks can be calculated as:

$$\theta = 2q(4n-1)(N-n+1) \quad (16)$$

The speed up ratio in this case can be computed as follows:

$$\eta = \frac{2q(4n-1)(N-n+1)}{(2q+1)(5N \log_2 N) + q(8N-n) + N} \quad (17)$$

Practical speed up ratio for searching short complex successive (n) data in a long complex-valued input vector (L) using complex-valued time delay neural networks is shown in Fig. 5. This has been performed by using a 700 MHz processor and MATLAB.

B) For a two dimensional input matrix

Multiplication of (n^2) complex-valued weights by (n^2) real inputs requires $(6n^2)$ real operations. This produces (n^2) real numbers and (n^2) imaginary numbers. The addition of these numbers requires $(2n^2-2)$ real operations. Therefore, the number of computation steps required by conventional neural networks can be calculated as:

$$\theta = 2q(4n^2-1)(N-n+1)^2 \quad (18)$$

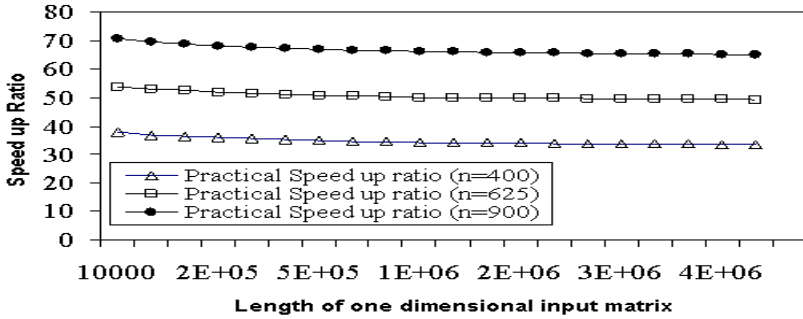


Fig. 5. Practical speed up ratio when using FTDNNs in case of one dimensional complex-valued input matrix and complex-valued weights

The speed up ratio in this case can be computed as follows:

$$\eta = \frac{2q(4n^2 - 1)(N - n + 1)^2}{(2q + 1)(5N^2 \log_2 N^2) + q(8N^2 - n^2) + N} \tag{19}$$

Practical speed up ratio for detecting (nxn) complex-valued submatrix in a large complex-valued matrix (NxN) using complex-valued neural networks is shown in Fig. 6. This has been performed by using a 700 MHz processor and MATLAB.

An interesting point is that the memory capacity is reduced when using FTDNN. This is because the number of variables is reduced compared with CTDNN.

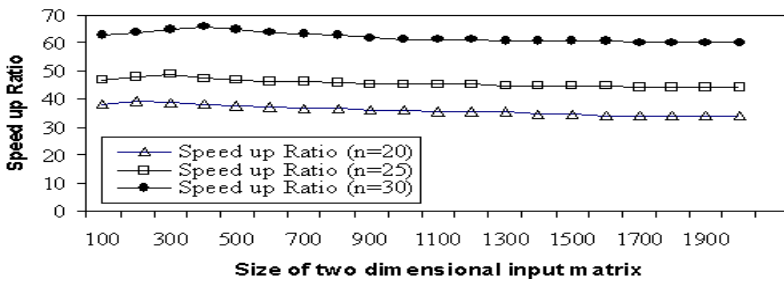


Fig. 6. Practical speed up ratio when using FTDNNs in case of two dimensional complex-valued input matrix and complex-valued weights

5 Conclusion

A fast neural algorithm for detecting important information in DNA sequence has been presented. Such strategy has been realized by using our design for FTDNNs. Theoretical computations have shown that FTDNNs require fewer computation steps than conventional ones. This has been achieved by applying cross correlation in the frequency domain between the input data and the weights of neural networks. Simulation results have confirmed this proof by using MATLAB. The proposed algorithm can be applied

successfully for solving many other bioinformatics problems like protein structure prediction, RNA structure prediction, promoter region identification, etc.

References

- [1] El-Bakry, H.M.: New Faster Normalized Neural Networks for Sub-Matrix Detection using Cross Correlation in the Frequency Domain and Matrix Decomposition. *Applied Soft Computing Journal* 8(2), 1131–1149 (2008)
- [2] El-Bakry, H.M., Zhao, Q.: Fast Pattern Detection Using Normalized Neural Networks and Cross Correlation in the Frequency Domain. *EURASIP Journal on Applied Signal Processing, Special Issue on Advances in Intelligent Vision Systems: Methods and Applications—Part I 2005*(13), 2054–2060 (2005)
- [3] El-Bakry, H.M., Zhao, Q.: A Fast Neural Algorithm for Serial Code Detection in a Stream of Sequential Data. *International Journal of Information Technology* 2(1), 71–90 (2005)
- [4] Hirose, A.: *Complex-Valued Neural Networks Theories and Applications*. Series on innovative Intelligence, vol. 5 (November 2003)
- [5] El-Bakry, H.M.: Face detection using fast neural networks and image decomposition. *Neurocomputing Journal* 48, 1039–1046 (2002)
- [6] El-Bakry, H.M.: Human Iris Detection Using Fast Cooperative Modular Neural Nets and Image Decomposition. *Machine Graphics & Vision Journal (MG&V)* 11(4), 498–512 (2002)
- [7] El-Bakry, H.M.: Automatic Human Face Recognition Using Modular Neural Networks. *Machine Graphics & Vision Journal (MG&V)* 10(1), 47–73 (2001)
- [8] Jankowski, S., Lozowski, A., Zurada, M.: Complex-valued Multistate Neural Associative Memory. *IEEE Trans. on Neural Networks* 7, 1491–1496 (1996)
- [9] El-Bakry, H.M.: New Fast Principal Component Analysis for Face Detection. *Journal of Advanced Computational Intelligence and Intelligent Informatics* 11(2), 195–201 (2007)
- [10] El-Bakry, H.M., Zhao, Q.: Speeding-up Normalized Neural Networks For Face/Object Detection. *Machine Graphics & Vision Journal (MG&V)* 14(1), 29–59 (2005)
- [11] Cooley, J.W., Tukey, J.W.: An algorithm for the machine calculation of complex Fourier series. *Math. Comput.* 19, 297–301 (1965)
- [12] Klette, R., Zamperon: *Handbook of image processing operators*. John Wiley & Sons Ltd., Chichester (1996)
- [13] Snyder, E.E., Stormo, G.D.: Identification of Protein Coding Regions In Genomic DNA. *ICCS Transactions* (2002)
- [14] Audic, S., Claverie, J.-M.: Self-identification of protein-coding regions in microbial genomes. *Structural and Genetic Information Laboratory, Centre National de la Recherche Scientifique-EP* 91 (2002)
- [15] Fickett, J.: Recognition of protein coding regions in DNA sequences. *Nucleic Acids Res.* 10, 5303–5318 (1982)
- [16] Farber, R., Lapedes, A., Sirotkin, K.: Determination of eukaryotic protein coding regions using neural networks and information theory. *J. Mol. Biol.* 226, 471–479 (1992)

Consistency-Based Feature Selection

Kilho Shin and Xian Ming Xu

Carnegie Mellon CyLab, Japan
 yshin@cmuj.jp

Abstract. Feature selection, the job to select features relevant to classification, is a central problem of machine learning. Inconsistency rate is known as an effective measure to evaluate consistency (relevance) of feature subsets, and INTERACT, a state-of-the-art feature selection algorithm, takes advantage of it. In this paper, we shows that inconsistency rate is not the unique measure of consistency by introducing two new consistency measures, and also, show that INTERACT has the important deficiency that it fails for particular types of probability distributions. To fix the deficiency, we propose two new algorithms, which have flexibility of taking advantage of any of the new measures as well as inconsistency rate. Furthermore, through experiments, we compare the three consistency measures, and prove effectiveness of the new algorithms.

1 Introduction

In machine learning problems, data are represented as vectors (f_1, \dots, f_L, c) : f_i 's are values for features F_i 's, and c is a class label. The objective of classifier algorithms is to guess a class label c given a feature vector (f_1, \dots, f_L) , and large L definitely does harm to the performance of the classifier algorithms in terms of accuracy and efficiency. Feature selection, therefore, aims to reduce L by eliminating those features that are irrelevant to classification. Many feature selection algorithms in the literature evaluate relevance to classification per individual feature F . Such algorithms employ measures such as *mutual information* $I(F; C)$ ([1]) and *symmetric uncertainty* $SU(F; C)$ ([2,3,4]) over given sample datasets, and eliminate features with lower evaluation, where the variable C represents class labels. This approach, however, has the problem of ignoring interaction among relevant features.

For example, let F_1, F_2, G_1 and G_2 be binary features such that $\Pr[F_1 = f_1, F_2 = f_2, G_1 = g_1, G_2 = g_2]$ for $f_1, f_2, g_1, g_2 \in \{0, 1\}$ are defined as depicted in the right. When we let $C =$

		g_1		0		1		1		
f_1	$f_2 \setminus g_2$	0	$\frac{\epsilon}{12}$	0	$\frac{\epsilon}{12}$	1	$\frac{\epsilon}{12}$	1	$\frac{\epsilon}{12}$	Sum
0	0	$\frac{1}{16} + \frac{\epsilon}{12}$	$\frac{1}{16} + \frac{\epsilon}{12}$	$\frac{1}{16} + \frac{\epsilon}{12}$	$\frac{1}{16} + \frac{\epsilon}{12}$	$\frac{1}{16} - \frac{\epsilon}{4}$	$\frac{1}{16} - \frac{\epsilon}{4}$	$\frac{1}{16} + \frac{\epsilon}{4}$	$\frac{1}{16} + \frac{\epsilon}{4}$	$\frac{1}{4}$
0	1	$\frac{1}{16} - \frac{\epsilon}{12}$	$\frac{1}{16} - \frac{\epsilon}{12}$	$\frac{1}{16} - \frac{\epsilon}{12}$	$\frac{1}{16} - \frac{\epsilon}{12}$	$\frac{1}{16} + \frac{\epsilon}{4}$	$\frac{1}{16} + \frac{\epsilon}{4}$	$\frac{1}{16} - \frac{\epsilon}{4}$	$\frac{1}{16} - \frac{\epsilon}{4}$	$\frac{1}{4}$
1	0	$\frac{1}{16} - \frac{\epsilon}{12}$	$\frac{1}{16} - \frac{\epsilon}{12}$	$\frac{1}{16} - \frac{\epsilon}{12}$	$\frac{1}{16} - \frac{\epsilon}{12}$	$\frac{1}{16} + \frac{\epsilon}{4}$	$\frac{1}{16} + \frac{\epsilon}{4}$	$\frac{1}{16} - \frac{\epsilon}{4}$	$\frac{1}{16} - \frac{\epsilon}{4}$	$\frac{1}{4}$
1	1	$\frac{1}{16} + \frac{\epsilon}{12}$	$\frac{1}{16} + \frac{\epsilon}{12}$	$\frac{1}{16} + \frac{\epsilon}{12}$	$\frac{1}{16} + \frac{\epsilon}{12}$	$\frac{1}{16} - \frac{\epsilon}{4}$	$\frac{1}{16} - \frac{\epsilon}{4}$	$\frac{1}{16} + \frac{\epsilon}{4}$	$\frac{1}{16} + \frac{\epsilon}{4}$	$\frac{1}{4}$
$C = 0$		$\frac{1}{8} + \frac{\epsilon}{6}$	$\frac{1}{8} + \frac{\epsilon}{6}$	$\frac{1}{8} + \frac{\epsilon}{6}$	$\frac{1}{8} + \frac{\epsilon}{6}$	$\frac{1}{8} - \frac{\epsilon}{2}$	$\frac{1}{8} - \frac{\epsilon}{2}$	$\frac{1}{8} + \frac{\epsilon}{2}$	$\frac{1}{8} + \frac{\epsilon}{2}$	$\frac{1}{2}$
$C = 1$		$\frac{1}{8} - \frac{\epsilon}{6}$	$\frac{1}{8} - \frac{\epsilon}{6}$	$\frac{1}{8} - \frac{\epsilon}{6}$	$\frac{1}{8} - \frac{\epsilon}{6}$	$\frac{1}{8} + \frac{\epsilon}{2}$	$\frac{1}{8} + \frac{\epsilon}{2}$	$\frac{1}{8} - \frac{\epsilon}{2}$	$\frac{1}{8} - \frac{\epsilon}{2}$	$\frac{1}{2}$
Sum		$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	1

$F_1 \oplus F_2, I(F_i; C) = SU(F_i; C) = 0$ holds. Thus, F_1 and F_2 interact each other to determine C , but each has no relevance to C . On the other hand, the relevance of G_i to C is calculated as $I(G_i; C) = \frac{3+4\epsilon}{6} \cdot \ln \frac{3+4\epsilon}{3} + \frac{3-4\epsilon}{6} \cdot \ln \frac{3-4\epsilon}{3} > 0$ and $SU(G_i; C) = 2 \cdot \frac{I(G_i; C)}{H(G_i)+H(C)} = \frac{I(G_i; C)}{\ln 2} > 0$. Hence, those feature selection algorithms of the aforesaid type definitely select (G_1, G_2) rather than (F_1, F_2) , and the predictive accuracy of any classifier performing based on (G_1, G_2) cannot exceed $3 \cdot (\frac{1}{8} + \frac{\epsilon}{6}) + (\frac{1}{8} + \frac{\epsilon}{2}) = \frac{1}{2} + \epsilon$.

The consistency-based approach (e.g. [5]) can be a solution to this problem. We let \mathcal{E} be a finite set of samples (training data, examples) with respect to a feature set $\mathcal{F} = \{F_1, \dots, F_L\}$ and a class variable C .

Definition 1. A subset $\tilde{\mathcal{F}} \subseteq \mathcal{F}$ is said to be \mathcal{E} -consistent or simply consistent to C , if, and only if, $c = c'$ holds for arbitrary $((f_1, \dots, f_L, c), (f'_1, \dots, f'_L, c')) \in \mathcal{E} \times \mathcal{E}$ such that $f_i = f'_i$ for all of $F_i \in \tilde{\mathcal{F}}$. We also say that \mathcal{F} determines C .

To the best of our knowledge, FOCUS ([6]) is the first instance of the consistency-based filters. On input of \mathcal{F} and \mathcal{E} , FOCUS selects the smallest \mathcal{E} -consistent subset $\tilde{\mathcal{F}} \subseteq \mathcal{F}$. A problem of FOCUS is that, if \mathcal{F} is not \mathcal{E} -consistent due to noises in \mathcal{E} , FOCUS fails. Zhao and Liu ([7]) approached this problem by proposing inconsistency rate, a measure to evaluate the extent to which $\tilde{\mathcal{F}}$ is consistent to C , and INTERACT, a state-of-the-art consistency-based algorithm.

This paper is development of the direction proposed by Zhao and Liu ([7]), and its contribution is summarized as follows.

1. In addition to inconsistency rate, we derive two measures of consistency from well-known information theoretic measures, and clarify quantitative relation among the measures including inconsistency rate (Section [2]).
2. We describe an important deficiency of the design of INTERACT, and show that it fails for particular types of probability distributions (Section [3]).
3. To fix the deficiency, we propose two new algorithms (Section [4]).
4. Through experiments, we verify effectiveness of the proposed algorithms, and compare the three measures of consistency (Section [5]).

2 Measures of Consistency

2.1 Inconsistency Rate and Consistency Measures

We generalize the definition of inconsistency rate by Zhao and Liu ([7]) only for finite datasets so that it is applicable to arbitrary probability distributions.

Definition 2. \mathcal{F}^* denotes the random variable that represents the joint distribution of the features in \mathcal{F} as random variables, and $R(\mathcal{F}^*)$ and $R(C)$ do the ranges of \mathcal{F}^* and a class variable C . Inconsistency rate $ICR(\mathcal{F}; C)$ is defined by:

$$ICR(\mathcal{F}; C) = 1 - \sum_{f \in R(\mathcal{F}^*)} \max_{c \in R(C)} \Pr[\mathcal{F}^* = f, C = c]$$

The following properties of $ICR(\mathcal{F}; C)$ are important (see also [2,2]).

Proposition 1

1. (Determinacy) $ICR(\mathcal{F}; C) = 0$, if, and only if, \mathcal{F} determines C .
2. (Monotonicity) $ICR(\mathcal{F}; C) \leq ICR(\mathcal{G}; C)$, if $\mathcal{F} \supset \mathcal{G}$.
3. (Maximum) $ICR(\mathcal{F}; C) \leq \frac{n-1}{n}$, if C takes n class labels.

2.2 New Consistency Measures

The determinacy and monotonicity properties shown in Proposition 1 are to be viewed as fundamental requirements for a consistency measure $\mu(\mathcal{F}; C)$.

Determinacy: A particular value of $\mu(\mathcal{F}; C)$ should faithfully correspond to the property that \mathcal{F} is consistent to C .

Monotonicity: If $\mathcal{F} \supset \mathcal{G}$, \mathcal{F} is more consistent to C than \mathcal{G} is. Hence, $\mu(\mathcal{F}; C)$ should be monotonous according to inclusive relation of feature sets.

We introduce two consistency measures $H(C|\mathcal{F}^*)$ and $\overline{SU}(\mathcal{F}; C)$ by

$$H(C|\mathcal{F}^*) = H(C) - I(\mathcal{F}^*; C) \text{ and}$$

$$\overline{SU}(\mathcal{F}^*; C) = \frac{2 \cdot H(C)}{H(\mathcal{F}^*) + H(C)} - SU(\mathcal{F}^*; C) = \frac{2 \cdot H(C|\mathcal{F}^*)}{H(\mathcal{F}^*) + H(C)}.$$

$H(C|\mathcal{F}^*)$ and $\overline{SU}(\mathcal{F}; C)$ are so defined that they align with $ICR(\mathcal{F}; C)$, and share the determinacy and monotonicity properties of the following form.

Proposition 2. *When $\mu(\mathcal{F}; C)$ denotes either $H(C|\mathcal{F}^*)$ or $\overline{SU}(\mathcal{F}^*; C)$, the following properties hold.*

Determinacy: $\mu(\mathcal{F}; C) = 0$, if, and only if, \mathcal{F} determines C .

Monotonicity: $\mu(\mathcal{F}; C) \leq \mu(\mathcal{G}; C)$, if $\mathcal{F} \supset \mathcal{G}$.

$ICR(\mathcal{F}; C)$, $H(C|\mathcal{F}^*)$ and $\overline{SU}(\mathcal{F}; C)$ have the following characteristics as consistency measures, respectively.

$ICR(\mathcal{F}; C)$ for the population is identical to the theoretical upper limit of the predictive accuracy of classifier algorithms.

$H(C|\mathcal{F}^*) = \sum_{f \in R(\mathcal{F}^*)} \Pr[\mathcal{F}^* = f] \cdot H(C | \mathcal{F}^* = f)$ is the average of $H(C | \mathcal{F}^* = f)$, which represents uncertainty associated with C when $\mathcal{F}^* = f$.

$\overline{SU}(\mathcal{F}; C)$ evaluates uncertainty of features in \mathcal{F} in addition. In fact, for the same $H(C|\mathcal{F}^*)$ and $H(C)$, $\overline{SU}(\mathcal{F}; C)$ decreases as $H(\mathcal{F}^*)$ increases.

2.3 Quantitative Relation among the Consistency Measures

Since $\overline{SU}(\mathcal{F}; C)$ is mathematically derived from $H(C|\mathcal{F}^*)$, its quantitative relation to $H(C|\mathcal{F}^*)$ is apparent. On the other hand, the relation between $H(C|\mathcal{F}^*)$ and $ICR(\mathcal{F}; C)$ is given by Theorem 1.

Theorem 1. Let \mathcal{F} and C be a feature set and a class variable such that the ranges of \mathcal{F}^* and C are $R(\mathcal{F}^*) = \{f_1, \dots, f_m\}$ and $R(C) = \{c_1, \dots, c_n\}$ for $m \geq 2$ and $n \geq 2$. Given r such that $ICR(\mathcal{F}; C) = 1 - r$, the following formulas give the maximum and the minimum of $H(C|\mathcal{F}^*)$.

$$\begin{aligned} \max_{ICR(\mathcal{F}; C)=1-r} H(C|\mathcal{F}^*) &= -r \ln r - (1 - r) \ln \frac{1 - r}{n - 1} \\ \min_{ICR(\mathcal{F}; C)=1-r} H(C|\mathcal{F}^*) &= \left(\left\lfloor \frac{1}{r} \right\rfloor + 1 \right) \left(1 - r \left\lfloor \frac{1}{r} \right\rfloor \right) \ln \left(\left\lfloor \frac{1}{r} \right\rfloor + 1 \right) \\ &\quad + \left\lfloor \frac{1}{r} \right\rfloor \left(r \left(\left\lfloor \frac{1}{r} \right\rfloor + 1 \right) - 1 \right) \ln \left\lfloor \frac{1}{r} \right\rfloor \end{aligned}$$

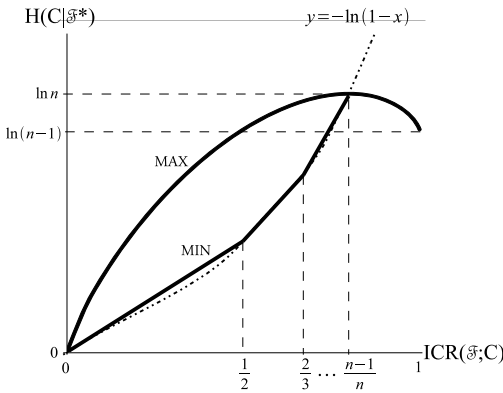


Fig. 1. $ICR(\mathcal{F}; C)$ and $H(C|\mathcal{F}^*)$

Although Theorem 1 indicates overall correlation between $H(C|\mathcal{F}^*)$ and $ICR(\mathcal{F}; C)$ (as depicted by Figure 1), the correlation is weak in a neighborhood of the origin. In fact, the ratio between the maximum and the minimum of $H(C|\mathcal{F}^*)$ diverges to infinity as $ICR(\mathcal{F}; C)$ approaches 0. This implies that a feature selection algorithm may show different performance when it employs $H(C|\mathcal{F}^*)$ and $ICR(\mathcal{F}; C)$, since the algorithm searches answers within a range of small consistency measures.

2.4 A Proof of Theorem 1 (a sketch)

The maximum property is proved by Lagrange’s method of undetermined multipliers. On the other hand, by Lemma 1, proving the minimum property is reduced to solving the optimization problem defined below, where $p_i = \Pr[\mathcal{F}^* = f_i]$ and $r_i = \max_{j=1, \dots, n} \Pr[C = c_j | \mathcal{F}^* = f_i]$.

$$\begin{aligned} \text{Minimize} \quad & \sum_{i=1}^m -p_i \left[\left\lfloor \frac{1}{r_i} \right\rfloor r_i \ln r_i + \left(1 - \left\lfloor \frac{1}{r_i} \right\rfloor r_i \right) \ln \left(1 - \left\lfloor \frac{1}{r_i} \right\rfloor r_i \right) \right] \\ \text{Subject to} \quad & \sum_{i=1}^m p_i = 1, \quad \sum_{i=1}^m p_i r_i = r, \quad p_i > 0, \quad \frac{1}{n} \leq r_i \leq 1 \end{aligned}$$

Lemma 1. Let a and b be positive constants with $\frac{a}{k} \leq b \leq a$. When the variables x_1, \dots, x_k are subject to $\sum_{i=1}^k x_i = a$ and $0 \leq x_i \leq b$, we have

$$\min \left(- \sum_{i=1}^k x_i \ln x_i \right) = - \left\lfloor \frac{a}{b} \right\rfloor b \ln b - \left(a - \left\lfloor \frac{a}{b} \right\rfloor b \right) \ln \left(a - \left\lfloor \frac{a}{b} \right\rfloor b \right).$$

3 A Deficiency of the Design of INTERACT

Algorithm: INTERACT (7)
 INPUT: A feature set \mathcal{F} , an example set \mathcal{E} , a threshold δ
 OUTPUT: A feature subset $\tilde{\mathcal{F}}$
 STEPS:
 Let $\tilde{\mathcal{F}} = \mathcal{F}$.
 Order the features F in $\tilde{\mathcal{F}}$ in incremental order of $SU(F; C)$.
 For each $F \in \tilde{\mathcal{F}}$ from the first to the end.
 If $CC(F, \tilde{\mathcal{F}}) \leq \delta$, let $\tilde{\mathcal{F}} = \tilde{\mathcal{F}} \setminus \{F\}$.
 End For.

Fig. 2. The algorithm of INTERACT

In designing INTERACT, Zhao and Liu (7) defined *consistency contribution* by $CC(F, \tilde{\mathcal{F}}) = ICR(\tilde{\mathcal{F}} \setminus \{F\}; C) - ICR(\tilde{\mathcal{F}}; C)$ to evaluate the contribution of an individual feature $F \in \tilde{\mathcal{F}}$ to $ICR(\tilde{\mathcal{F}}; C)$. As Figure 2 depicts, INTERACT sets $\tilde{\mathcal{F}} = \mathcal{F}$ as the initial value, examines a feature $F \in \mathcal{F}$ one by one in the incremental order of $SU(F; C)$, and eliminates F from $\tilde{\mathcal{F}}$, if $CC(F, \tilde{\mathcal{F}}) \leq \delta$ for a given threshold δ .

A problem of INTERACT is that, even if the threshold δ is set small, $ICR(\tilde{\mathcal{F}}; C)$ for its output $\tilde{\mathcal{F}}$ can be large. Let $\tilde{\mathcal{F}}_0 \supsetneq \tilde{\mathcal{F}}_1 \supsetneq \dots \supsetneq \tilde{\mathcal{F}}_L$ be a history of the values of the variable $\tilde{\mathcal{F}}$ held by INTERACT. Although $ICR(\tilde{\mathcal{F}}_{k+1}; C) - ICR(\tilde{\mathcal{F}}_k; C) \leq \delta$, $ICR(\tilde{\mathcal{F}}_L; C)$ can be large even for small δ . Example 1 presents an extreme case, where INTERACT eliminates all the features, and there certainly exist many different probability distributions for which INTERACT fails.

Example 1. We let F_i and C be boolean, and let $\mathcal{F}^* = b$ represent the event of $F_i = b_i$ for $i = 1, \dots, L$, where b_i denotes the i -th least significant bit of an integer $b \in [0, 2^L)$. Then, for a rational number δ' such that $\delta' \leq 1/L$, we let \mathcal{E} be a finite sample set with the probability distribution defined as follows.

$$\Pr[\mathcal{F}^* = b, C = c] = \begin{cases} 1 - L\delta', & \text{if } b = 0 \text{ and } c = 0, \\ \delta', & \text{if } b = 2^{i-1} \text{ for } \exists i \in \{1, \dots, L\} \text{ and } c = 1, \\ 0, & \text{otherwise.} \end{cases}$$

Since none of F_i is particular to the others, we can assume that INTERACT evaluates F_k from $k = L$ to 1. We let $\tilde{\mathcal{F}}_k = \{F_1, \dots, F_{L-k}\}$. Then, for $b' \in [0, 2^{L-k})$, $\Pr[\tilde{\mathcal{F}}_k^* = b', C = c]$ proves to be: $1 - L\delta'$, if $b' = 0$ and $c = 0$; $k\delta'$, if $b' = 0$ and $c = 1$; δ' , if $b' = 2^{i-1}$ for some $i \in \{1, \dots, L - k\}$ and $c = 1$; and 0, otherwise. Thus, the following equalities hold, and hence, INTERACT cannot help failing by outputting $\tilde{\mathcal{F}}_L = \emptyset$ for $\delta \leq \delta'$.

$$ICR(\tilde{\mathcal{F}}_k; C) = \min\{1 - L\delta', k\delta'\}$$

$$CC(F_{L-k}, \tilde{\mathcal{F}}_k) = \begin{cases} \delta', & \text{if } k = 0, \dots, \lfloor \frac{1}{\delta'} \rfloor - L - 1, \\ 1 - \delta' \lfloor \frac{1}{\delta'} \rfloor, & \text{if } k = \lfloor \frac{1}{\delta'} \rfloor - L, \\ 0, & \text{if } k = \lfloor \frac{1}{\delta'} \rfloor - L + 1, \dots, L - 1. \end{cases}$$

4 Consistency-Constrained Feature Selection Algorithms

Algorithm: Linear CC (LCC)
INPUT: A measurement function μ ,
an ordered feature set \mathcal{F} ,
an example set \mathcal{E} , a threshold δ
OUTPUT: A minimal subset $\tilde{\mathcal{F}} \subseteq \mathcal{F}$ such that
 $\mu(\tilde{\mathcal{F}}; C) \leq \delta$.
STEPS:
Let $\tilde{\mathcal{F}} = \mathcal{F}$.
If $\mu(\tilde{\mathcal{F}}; C) > \delta$, abort.
For each $F \in \mathcal{F}$ from the first to the end.
 If $\mu(\tilde{\mathcal{F}} \setminus \{F\}; C) \leq \delta$, let $\tilde{\mathcal{F}} = \tilde{\mathcal{F}} \setminus \{F\}$.
End For.

feature set \mathcal{F} (e.g. ordered in the incremental order of $SU(F; C)$). It first sets $\tilde{\mathcal{F}} = \mathcal{F}$, examines F in the given order, and eliminates F from $\tilde{\mathcal{F}}$, if $\mu(\tilde{\mathcal{F}} \setminus \{F\}; C) \leq \delta$. The resultant $\tilde{\mathcal{F}}$ is minimal (no $\mathcal{G} \subsetneq \tilde{\mathcal{F}}$ meets $\mu(\mathcal{G}; C) \leq \delta$).

Algorithm: Complete CC (CCC)
INPUT: A consistency measure function μ ,
a feature set \mathcal{F} ,
an example set \mathcal{E} , a threshold δ
OUTPUT: A smallest subset $\tilde{\mathcal{F}} \subseteq \mathcal{F}$ such that
 $\mu(\tilde{\mathcal{F}}; C) \leq \delta$.
STEPS:
Let $\tilde{\mathcal{F}} = \emptyset$.
For **size** = 1 to $\|\mathcal{F}\|$
 Find $\tilde{\mathcal{F}}$ with $\|\tilde{\mathcal{F}}\| = \mathbf{size}$ and
 the smallest $\delta' = \mu(\tilde{\mathcal{F}}; C)$.
 If $\delta' \leq \delta$, return $\tilde{\mathcal{F}}$.
End For.

In order to fix the aforesaid deficiency of INTERACT, we introduce two feature selection algorithms, namely, Linear Consistency-Constrained (LCC) and Complete Consistency-Constrained (CCC) algorithms.

Both take a measurement function μ as an input, and outputs a feature subset $\tilde{\mathcal{F}}$ such that $\mu(\tilde{\mathcal{F}}; C) \leq \delta$ for a threshold δ .

LCC receives an ordered feature set \mathcal{F} (e.g. ordered in the incremental order of $SU(F; C)$). It first sets $\tilde{\mathcal{F}} = \mathcal{F}$, examines F in the given order, and eliminates F from $\tilde{\mathcal{F}}$, if $\mu(\tilde{\mathcal{F}} \setminus \{F\}; C) \leq \delta$. The resultant $\tilde{\mathcal{F}}$ is minimal (no $\mathcal{G} \subsetneq \tilde{\mathcal{F}}$ meets $\mu(\mathcal{G}; C) \leq \delta$).

By contrast, CCC first sets a variable **size** to 1, and attempts to find $\tilde{\mathcal{F}}$ with $\|\tilde{\mathcal{F}}\| = \mathbf{size}$ that meets $\mu(\tilde{\mathcal{F}}; C) \leq \delta$. If it cannot find such $\tilde{\mathcal{F}}$, it increments **size** by one, and then continues the search. The resultant $\tilde{\mathcal{F}}$ is a feature subset with the smallest **size** such that $\mu(\tilde{\mathcal{F}}; C) \leq \delta$.

LCC evaluates $\mu(\tilde{\mathcal{F}}; C)$ for different subsets $\tilde{\mathcal{F}}$ in $\|\mathcal{F}\|$ times, while CCC does in $2^{\|\mathcal{F}\|}$ times in the worst case.

5 Experimental Results

We compare the algorithms and the consistency measures through experiments.

For the purpose, we defined four types of synthetic datasets as defined in Table 1, and generated 10 datasets at random per type. Each type includes binary *relevant* features F_i ($i = 1, \dots, k$), binary *irrelevant* features $G_{i,j}$ ($i, j = 1, \dots, \ell$) and a binary class variable C . Also, it is designed based on Example 1 so that a feature subset $\tilde{\mathcal{F}}$ is consistent to C , if, and only if, $\tilde{\mathcal{F}} \supseteq \{F_1, \dots, F_k\}$. Also, we selected four real datasets as shown in Table 2.

To each dataset, we applied five algorithms, *i.e.* INTERACT, CCC with $ICR(\tilde{\mathcal{F}}; C)$ and LCC with $ICR(\tilde{\mathcal{F}}; C)$, $H(C|\tilde{\mathcal{F}}^*)$ and $\overline{SU}(\tilde{\mathcal{F}}; C)$, changing the threshold δ , and evaluated their outputs $\tilde{\mathcal{F}}$ as shown in Table 3. In particular, for the synthetic datasets, we used the coverage rate R_c and the positively false rate R_p , and, we experimented with the real datasets using the classifier algorithms implemented in Weka ([10]).

5.1 Comparison of the Algorithms

Figure 3 depicts the comparison of INTERACT, LCC and CCC with $ICR(\tilde{\mathcal{F}}; C)$ for the synthetic datasets, and clearly exhibits inferiority of INTERACT.

R_c drops significantly more rapidly than the other two, and R_p for Type #4 increases more rapidly. This implies that INTERACT has only narrow latitude for the selection of δ , and hence, has weak tolerance to noises. Also, Figure 4 exhibits inferiority of INTERACT, when applied to the real datasets.

Not surprisingly, CCC couldn't return answers for the larger datasets (Type #4, Kr-vs-Kp and ADA) within the time allowance of 30 minutes, in contrast that LCC and INTERACT returned answers within 10 seconds per experiment.

5.2 Comparison of the Consistency Measures

For comparison of the consistency measures, we applied LCC to the same datasets changing the measures out of $H(C|\tilde{\mathcal{F}}^*)$, $\overline{SU}(\tilde{\mathcal{F}}; C)$ and $ICR(\tilde{\mathcal{F}}; C)$.

Table 1. Types of synthetic datasets

Relevant features: F_i ($i = 1, \dots, k$)
 Irrelevant features: $G_{i,j}$ ($i, j = 1, \dots, \ell$)

Parameters	#1	#2	#3	#4
$k =$	5	5	5	10
$\ell =$	3	3	3	4
Size of dataset =	100	100	100	1000
$\Pr[\mathcal{F}^* = 0, C = 0]$	0.85	0.55	0.55	0.725
$\Pr[\mathcal{F}^* = 2^{i-1}, C = 1]$	0.03	0.09	0.03i	0.005i
$\Pr[G_{i,j} = 0 C = 0] \approx$	0.25i	0.25i	0.25i	0.2i
$\Pr[G_{i,j} = 0 C = 1] \approx$	0.25j	0.25j	0.25j	0.2j

Table 2. Real datasets

Dataset Name	Source	# of Examples	# of Features	# of Labels
Wine	[8]	178	13	3
Zoo	[8]	101	16	7
Kr-vs-Kp	[8]	3196	36	2
ADA	[9]	4147	48	2

Table 3. Evaluation methods

Comparison	Dataset	Measures of evaluation
Algorithms/ Measures	Synthetic	R_c , the ratio of the number of F_i 's in $\tilde{\mathcal{F}}$ to k , and R_p , the ratio of the number of G_i 's in $\tilde{\mathcal{F}}$ to $ \tilde{\mathcal{F}} $.
Algorithms	Real	$ICR(\tilde{\mathcal{F}}; C)$ and "Percent Correct" with the C4.5 classifier.
Measures	Real	"Percent Correct" with the C4.5 and LogitBoost classifiers.

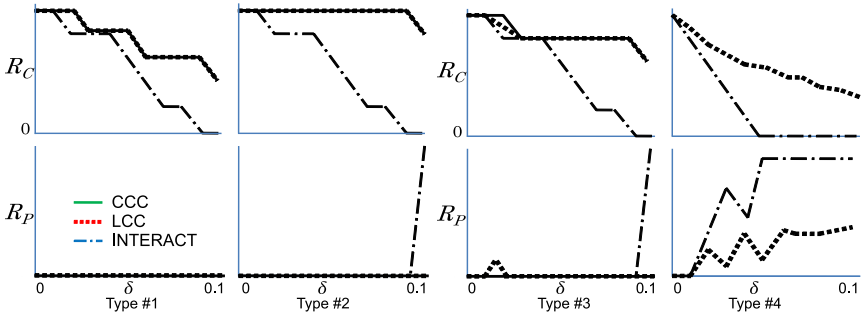


Fig. 3. Comparison of algorithms with the synthetic datasets

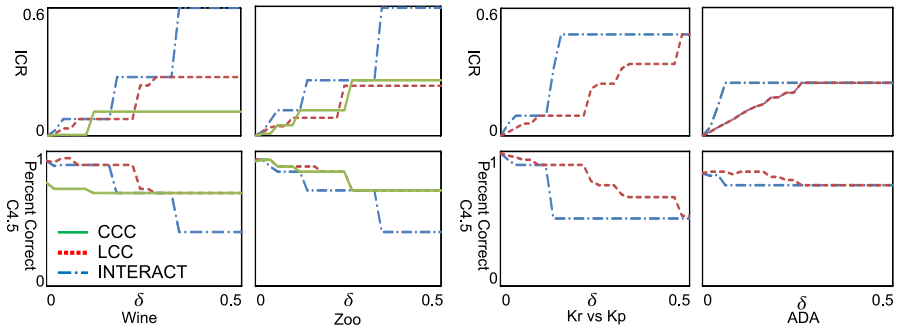


Fig. 4. Comparison of algorithms with the real datasets

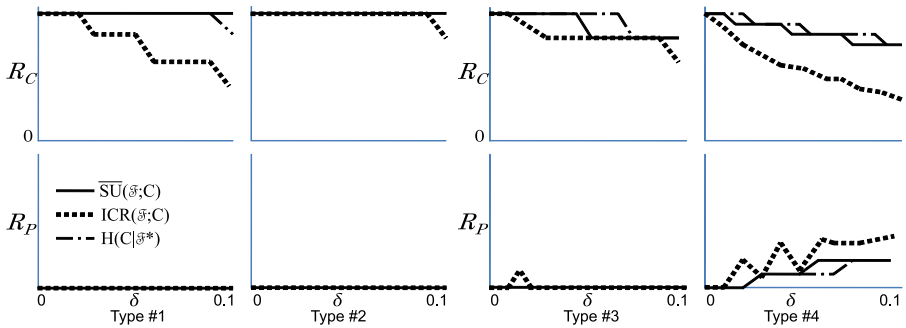


Fig. 5. Comparison of measures with the synthetic datasets

In the experiments, we took difference in scale between the measures into account by adjusting the threshold δ . For example, when $H(C|\tilde{\mathcal{F}}^*)$ was applied to the datasets with the binary class variable, we used $2\delta \ln 2$ instead of using δ as it is (Theorem [11](#)). Nevertheless, we should not take the difference of the

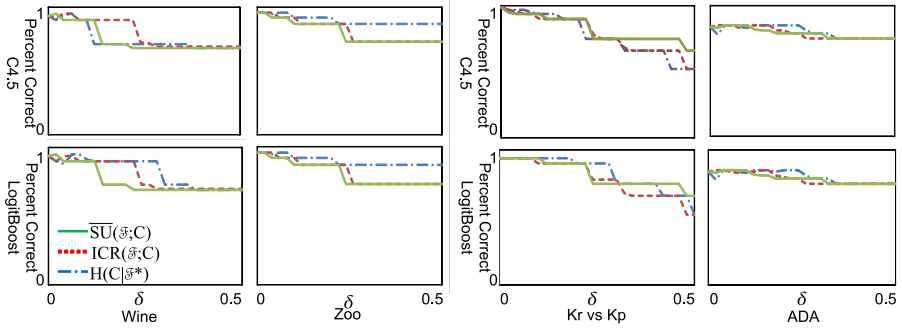


Fig. 6. Comparison of measures with the real datasets

results in the horizontal direction in Figure 5 and 6 too significant. When we keep this in mind, we see that $H(C|\tilde{\mathcal{F}}^*)$ exhibits the best results, and $ICR(\tilde{\mathcal{F}};C)$ exhibits the worst results for the datasets of Type #1, #4 and Zoo. For the other datasets, the measures shows compatible results.

References

1. Peng, H., Long, F., Ding, C.: Feature selection based on mutual information: Criteria of max-dependency, max-relevance and min-redundancy. *IEEE Transaction on Pattern Analysis and Machine Intelligence* 27(8) (2005)
2. Yu, L., Liu, H.: Feature selection for high-dimensional data: a fast correlation-based filter solution. In: *International Conference of Machine Learning* (2003)
3. Biesiada, J., Duch, W.: Feature selection for high-dimensional data – a Kolmogorov-Smirnov correlation-based filter. *Advances in Soft Computing* 30, 95–103 (2005)
4. Biesiada, J., Duch, W.: Feature selection for high-dimensional data – a Pearson redundancy based filter. *Advances in Soft Computing* 45, 242–249 (2008)
5. Dash, M., Liu, H.: Consistency-based search in feature selection. *Artificial Intelligence* 151, 155–176 (2003)
6. Almuallim, H., Dietterich, T.G.: Learning boolean concepts in the presence of many irrelevant features. *Artificial Intelligence* 69(1-2) (1994)
7. Zhao, Z., Liu, H.: Searching for interacting features. In: *Proceedings of International Joint Conference on Artificial Intelligence*, pp. 1156–1161 (2007)
8. Blake, C.S., Merz, C.J.: UCI repository of machine learning databases. Technical report. University of California, Irvine (1998)
9. *IEEEWorld Congress on Computational Intelligence. Performance prediction challenge* (2006), <http://www.modelselect.inf.ethz.ch/>
10. Witten, J.H., Frank, E.: *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers, San Francisco (2005)

Asynchronous Situated Coevolution and Embryonic Reproduction as a Means to Autonomously Coordinate Robot Teams

Abraham Prieto, Francisco Bellas, Andres Faina, and Richard J. Duro

Grupo Integrado de Ingeniería,
Universidade da Coruña, Spain
{abprieto, fran, afaina, richard}@udc.es

Abstract. One of the main challenges in the operation of multirobot systems is to find ways for them to adapt to changing situations and even objectives without any type of central control. In this work we propose a real time coevolutionary strategy based on Embodied Evolution (EE) approaches that provides a means to achieve this end. The main inspiration for this approach comes from the field of artificial life combined with some of the notions on the distribution of utility functions as proposed by the multiagent systems literature. The solution has been tested on different real life problems involving robot teams. In particular, in this paper the work is aimed at the coordination of sets of robots for performing monitoring and surveillance operations such as the ones required on ship tanks and hulls. Nevertheless, the approach is general enough to be applied to many other tasks in several fields.

Keywords: Coevolution, Adaptation, Multirobot Systems, Coordination.

1 Introduction

Embodied evolution (EE), as proposed by Watson et al. [1], aims at establishing a completely distributed evolutionary algorithm embodied in physical robots. The main idea behind the general approach is that a large number of robots may be used for the evaluation stage of an evolutionary process that aims to obtain a given controller for a particular task. Basically the authors seek a robot population that evolves in a completely autonomous manner, without external intervention. The main ideas of this approach are that the evolutionary process must be decentralized and that the evaluations required for the instantiation of the evolutionary operators acting on a given individual must take place directly within the individual in an embodied and localized manner, preferably on the physical robot itself. This deviates fundamentally from other strategies proposed in the Evolutionary Robotics (ER) community [2][3][4], whereby there is usually a centralized evolutionary algorithm that uses information from all the robots in a simulation (or even in some cases in real robots) in order to perform selection, crossover and whatever other operators the evolutionary process requires. As the authors indicate in [1] this approach opens many possibilities for the creation of collective behaviors in distributed robotics. However they do not really explore how to implement these and present examples mainly based on single robot behaviors.

On the other hand, in the last decade or two there has been a lot of effort invested in the production of coordinated behaviors for robot teams and swarms both, from the point of view of the formalization of the problems [5] in order to produce hand crafted algorithms or controllers for particular tasks or concentrating on implementation issues [6][7][8]. The main drawback of many of these approaches is that they are particular to a task (i.e. foraging or flocking) often using homogeneous sets of robots/controllers and do not provide a general framework for obtaining collective behaviors that consider the fact that the structure of most real world problems does not allow for an easy decomposition into subproblems nor are they known beforehand. This is especially so in the case of complex dynamic problems where the environment or even the objectives change with time.

In this paper we consider the EE based approach inspired on the main features of some artificial life based distributed evolutionary approaches and we combine it with some concepts from the multiagent systems literature in order to provide a way to implement the objectives desired from the collective system through the implementation of energy and interaction based utility distribution schemes. In addition, as robot teams usually have a fixed number of individuals, we provide an Embryonic based reproduction mechanism that deals with the instabilities of other direct implementations of reproduction in EE such as the Probabilistic Genetic Transfer Algorithm [1].

2 Asynchronous Situated Coevolution

The algorithm we consider here and which is called Asynchronous Situated Coevolution (ASiCo) is inspired, on one hand, on the studies of complex systems in terms of the use of decentralized and asynchronous evolution as introduced in artificial life simulations. Unlike other bioinspired approaches such as genetic algorithms in which selection and evaluation of the individuals is carried out in a centralized manner at regular processing intervals based on an objective function, this type of evolution is situated. This means that all of the interactions among individuals of the population are local and depend on spatial and temporal coincidence of the individuals, which implies an intrinsic decentralization. However, this type of evolution has usually been employed to study how evolution works in an open ended manner and not really with an engineering objective in mind. This is the reason why here we take inspiration from multiagent systems and their studies of utility functions and their distribution among the individuals in order to structure the energy dynamics of the environment in order to guide evolution to the objectives sought.

Consequently, the drive of the algorithm is provided by the interactions among elements in the environment and not by a sequence of optimization steps or processes as in traditional evolutionary algorithms. Reproduction, creation of new individuals or their elimination is driven by events that may occur in the environment and that correspond to interactions among individuals in the population. It is the design of these interactions that provides the power to achieve the objectives we seek.

Fig. 1 displays a schematic representation of the algorithm's structure, divided into two different parts. On one hand (left block) we have the *evolution engine*, which is based on the interactions among elements in the environment. After the *creation of a random population*, the execution of the interaction events occurs in a continuous loop through the actions performed by the individuals following their control structure.

These actions lead first to updating the *energy* values of each individual, then, if some predefined criteria is met, *reproductive selection* events occur, and finally, those individuals whose energy levels dip below a given threshold or who meet any other elimination criteria are *eliminated*. On the other hand (right block of Fig. 1), we have the procedures that guide this evolution towards an *objective*. The *individual encoding* defines the behavior that can be generated, the *behavior/objective function* establishes the energy assignment and reproductive selection criteria, and the *embryonic reproduction* which allows a situated evolution, and that will be explained in detail in the next section. In addition, Fig. 1 shows the relations between these procedures and the processes carried out during evolution.

ASiCo is an interaction driven algorithm. Interactions are a set of rules that make the state of the elements and individuals in the environment change when some event occurs. This process is independent from the evolution of the population. Two elements are very relevant within ASiCo. On one hand we have the flow of energy, which represents the different rules that regulate energy variations and transmissions between the individuals and the environment and vice versa. On the other hand, reproductive selection is the set of rules that regulates the reproduction process. This selection process must be defined for each problem but it is usually based on spatial interactions together with some energetic criteria. When individuals are selected for reproduction (for instance, because they came together in space and had enough energy) their genotypes are somehow combined through a crossover operator and a new individual comes into being in the environment taking up some of the energy from its parents. There are also some rules that regulate when an individual dies off.

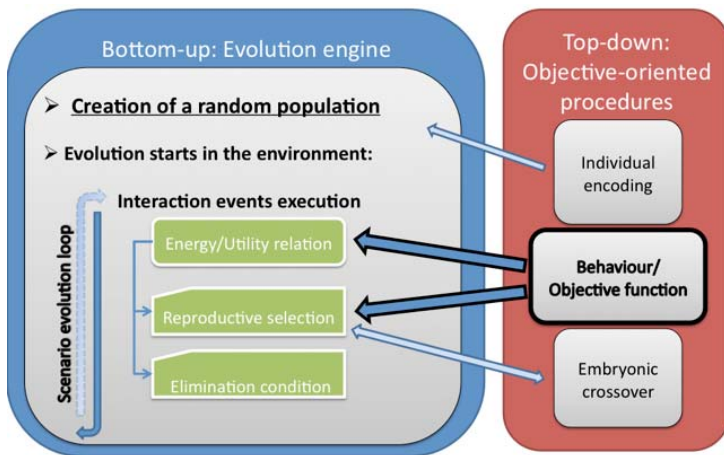


Fig. 1. Schematic representation of the interaction driven ASiCo

3 Embryo Based Reproduction

Regarding reproduction, this mechanism needs to be adapted to the objectives sought. In the work presented here we are interested in groups of real robots with a fixed

number of individuals. Consequently, we do not want real robots to “die” nor have the means to make robots “appear” in the environment and therefore the reproduction mechanism has to take this into account. The only way to do this, apart from continuously modifying on line the genetic make up of the robots (as is the case with PGTA [1]) with the consequent instabilities in their behaviors due to lack of evaluation time in the environment, is to synchronize death with birth. In fact, as the robots are preset, we can only work with their controllers and having a robot die and another one be born at the same time is basically an operation of changing the controller of a given robot. To allow for evolutionary pressure within this process, we have designed a reproduction mechanism for fixed size populations called Embryo Based Reproduction (EBR). The basic idea is that each agent, apart from its own evolvable parameters, carries within it, another set of parameters corresponding to its embryo with an associated pre-utility value that estimates the utility of the agent to be generated from the embryo. Initially, when a new agent is created, an embryo is generated for it as a mutation of the parent genotype with half of its energy. During the life of an agent, the embryo may be modified whenever the agent meets another one and evaluates it positively through an estimation of the pre-utility of the embryo after crossover by averaging the utility of the two parents. If this pre-utility is higher than the current one, crossover takes place and the new embryo substitutes the current one. Finally, when the parent dies because it ran out of energy or time or for whatever other reason, the embryo substitutes the parent, meaning, the control of the robotic unit is assumed by the embryo and a new embryo is generated within the robot.

Through this strategy, we ensure that the size of the population remains constant and that the process takes place in an asynchronous and decentralized way. In what follows we will present some examples of the results obtained using this approach.

4 Experiments and Results

An experiment was set up to ascertain the capabilities of Asynchronous Situated Co-evolution and EBR in terms of providing a mechanism for a fixed set of robots to be able to adapt in real time and establish coordination strategies through the evolution of their controllers for performing surveillance or monitoring tasks. Given the nature of the problems we are interested in, we have considered two scenarios. One where every area that needs to be monitored has the same importance or monitoring need and a second one where the risk of incidents increases differently for different areas when they are not monitored, which is a much more realistic situation albeit one that most authors ignore and which has a direct bearing on the efficiency of the coordinated strategy chosen.

In any case, for the sake of simplicity and clarity in the presentation of results, for the experiments presented here we have built an environment that is made up of cells and each cell is characterized by a level of risk that increases while no robot monitors it. The risk level goes to zero when it is monitored and starts to increase again until it is re-monitored by another robot. This set up allows us to measure the efficiency of the strategies employed by the robots in a very simple way. All we need to do is to add up the risk level in whatever area of the environment we are interested in.

To make the problem as complicated as possible (but still solvable) for the robots, we have infra-sensorized them. In fact, a robot can only detect on one hand how many robots are within the sensor reach (which is a constant for all the robots) in each of the four quadrants around it and, on the other, it detects if a collision has occurred. It has a small memory that allows it to remember its position n instants before. In terms of actions, all of the robots move at the same speed and the control system just provides a value for the angular velocity (rate of turn). In the first trial, the robots have no perception of the risk level at their location or of how it changes around them and, consequently, they cannot use this information in order to decide their actions.

The control system is based on a RBNN (Radial Basis Neural Network) whose parameters are encoded in the genotype of the robot together with the length of the memory. The inputs to this network are those corresponding to the number of robots around it within the four quadrants, the module and angle of the vector relating its position n instants before and its current position and the value of its collision sensor (whether a collision has occurred).

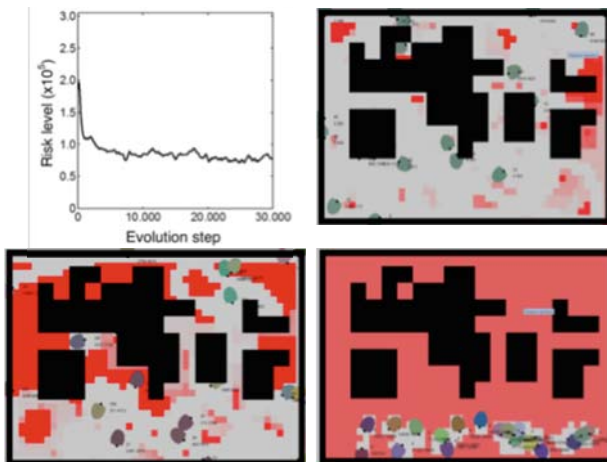


Fig. 2. Change in the risk level (top left) as well as three instants during evolution: 0 (top right), 1000 (bottom left) and 6000 (bottom right) steps. Black areas represent the obstacles. Intensity of red indicates risk level for a given area.

Several experiments were carried out with these elements in order to see how the algorithm behaved. The first one, based on the work in [9] considers the direct exploration of an environment where all of the cells behave similarly. Fig. 2 (top left) shows evolution of the risk level on an environment that is being controlled by a set of 20 robots as described above. The top right image of Fig. 2 displays the situation at the beginning of the process, where all the controllers are random. Red intensity represents risk level. The bottom left image corresponds to the state after around 1000 time steps in the environment and the bottom right one to the operation of the robot group after around 6000 steps. It is clear that the controllers of the robots have improved to the point of being able to obtain quite a low risk level (an average risk level

of around 50 units per cell which corresponds to every cell being explored every 450 time steps). This is a very adequate level if one takes into account that the robots, in order to coordinate their actions, are only aware of the number of agents in four quadrants around them and not of their exact position and they do not know the risk level of a given position. Consequently, the fact that the whole area is monitored efficiently is a consequence of an emergence of a coordination strategy forced by the global utility requirements.



Fig. 3. Initial behavior of a 20 robot system and after 8000 time steps for an environment with a central hot zone and neutral surroundings. Different genotypes lead to different robot colors.

Another interesting result is that the robots tend towards a homogeneous genotype in this task (different genotypes are shown as different colors for the robots). This seems reasonable given the fact that that the environment is relatively homogeneous in terms of requirements for the agents.

However, it would be nice to test the behavior of this type of approach in cases where different parts of the environment have different monitoring requirements. Thus, we considered a slightly different problem where different areas of the environment require surveillance or monitoring at different rates, that is, their risk levels are modified at different rates. This problem corresponds to one of the real problems we find in the shipbuilding industry where these strategies are being applied. When monitoring structural soundness and seeking faults, different types of elements require different monitoring, for instance, plates are less prone to damage than joints and plate-beam unions and thus must be explored at different rates.

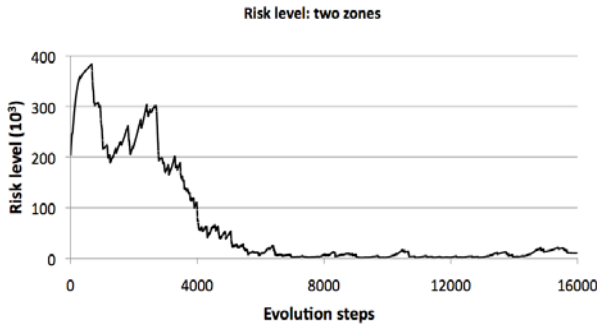


Fig. 4. Evolution of the risk level for the example of Fig. 3

For these experiments the scenario was modified to allow for different rates of risk evolution. In addition the robots were endowed with a sensor that detects transitions between areas (i.e. it knows when goes from a plate to a beam or vice versa). The first experiment carried out with different areas was using a single risk increasing area within a neutral area, that is, an area that does not need to be monitored. For simplicity we will differentiate area requirements using temperature analogies. Here case we have a hot area covering 15% of the scenario within a neutral area.

Fig. 3 displays the initial behavior and the genotypes (through robot color) of the robots and the behavior after 8000 time steps. Fig. 4 shows the evolution of the value of the risk for the whole environment. It is clear that this value reaches extremely low values after about 6000 time steps. In fact, the value achieved is very close to the theoretical minimum. This is a result of the production of an appropriate coordination strategy for the robots and the fact that there are so many robots to control in such a small area. Much in the same way as in the previous case, as all the robots have to work over the same area, they tend to have the same genotype, that is, the same controller, and thus become a homogeneous robot team.

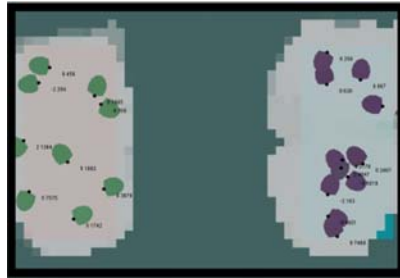


Fig. 5. Behavior of a 20 robot system for an environment with three different areas (hot, cold and neutral) after stability is achieved

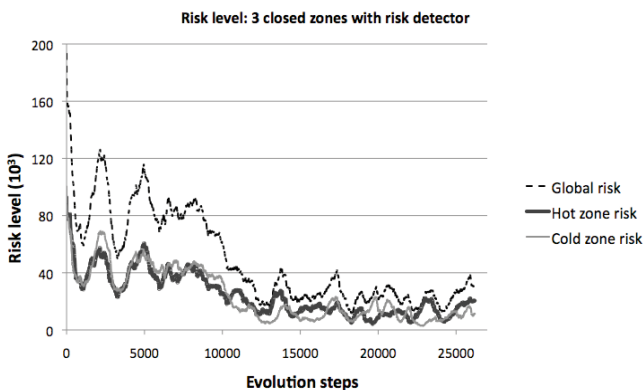


Fig. 6. Evolution of the risk level for the example of Fig. 5

To complicate things, now we establish a scenario with three different areas in terms of risk: a hot area, a cold area and a neutral area in between. In this case after reaching a stable behavior, as shown in Fig. 5, two types of robots are obtained. There is one group that specializes in surveillance of the cold area and another one that concentrates on the hot area. Obviously the robots are different in terms of controller (different genotype) as this makes the task simpler. Thus the robot population has evolved into a heterogeneous robot (or at least controller) group with species that specialize in different tasks. Fig. 6 displays the evolution of the global risk level as well as that of the two zones.

5 Conclusions

Asynchronous Situated Coevolution (ASiCo) together with an embryonic like delayed reproduction mechanism (EBR) provides a very simple approach for the introduction of real time evolution within robot coordination structures. It is a strategy that allows for the real time operation of a fixed number of robots that evolve their controllers in a decentralized and asynchronous manner when they interact in the environment optimizing very fast the use of the resources available to the team in order to achieve the task implicit in the global utility function distributed as energy based local utilities that determine the reproduction capabilities of the individuals. This approach has been explored here in a simulated environment that mimics the surveillance and monitoring tasks that need to be carried out in real environments such as ship fault detection and is now being implemented over real robots with promising results.

Acknowledgments. This work was partially funded by the MEC of Spain through projects DEP2006-56158-C03-02 and DPI2006-15346-C03-01.

References

1. Watson, R.A., Ficici, S.G., Pollack, J.B.: Embodied Evolution: Distributing an Evolutionary Algorithm in a Population of Robots. *Robotics and Autonomous Systems* 39(1), 1–18 (2002)
2. Harvey, I.: Artificial Evolution and Real Robots. In: Sugisaka, M. (ed.) *Proc. of the International Symposium on Artificial Life and Robotics*, Beppu, Japan, pp. 138–141 (1996)
3. Mataric, M.J., Cliff, D.: Challenges in Evolving Controllers for Physical Robots. *Journal of Robotics and Autonomous Systems* 19(1), 67–83 (1996)
4. Floreano, D., Mondada, F.: Evolutionary Neurocontrollers for Autonomous Mobile Robots. *Neural Networks* 11(7-8), 1461–1478 (1998)
5. Kolling, A., Carpin, S.: Multi-robot Surveillance: an Improved Algorithm for the GRAPH-CLEAR Problem. In: *Proc. 2008 IEEE International Conference on Robotics and Automation Pasadena, CA, USA, May 19-23*, pp. 2360–2365 (2008)
6. Burgard, W., Moors, M., Stachniss, C., Schneider, F.E.: Coordinated multi-robot exploration Robotics. *IEEE Transactions on Robotics* 21(3), 376–386 (2005)
7. Fox, D., Ko, J., Konolige, K., Limketkai, B., Schulz, D., Stewart, B.: Distributed Multirobot Exploration and Mapping. *Proc. of the IEEE* 94(7), 1325–1339 (2006)

8. Folgado, E., Rincón, M., Álvarez, J.R., Mira, J.: A Multi-robot Surveillance System Simulated in Gazebo. In: Mira, J., Álvarez, J.R. (eds.) IWINAC 2007. LNCS, vol. 4528, pp. 202–211. Springer, Heidelberg (2007)
9. Schut, M.C., Haasdijk, E., Prieto, A.: Is Situated Evolution an Alternative for Classical Evolution? Accepted at CEC 2009 (2009)

Learning Automata Based Intelligent Tutorial-like System

B. John Oommen¹ and M. Khaled Hashem²

¹ *Chancellor's Professor, Fellow: IEEE and Fellow: IAPR, School of Computer Science, Carleton University, Ottawa, Canada, Adjunct Professor with the University of Agder in Grimstad, Norway*

`oommen@scs.carleton.ca`

² *School of Computer Science, Carleton University, Ottawa, Canada*

`k_hashem@yahoo.com`

Abstract. The aim of this pioneering research¹ is to study, design, and implement systems that could tutor other sub-systems using techniques that traditional *real-life* Teachers use when they teach *real-life* Students. The research undertaken is a result of merging the fields of Intelligent Tutoring Systems (ITS) and Learning Automata (LA), and leads to a paradigm which we refer to as “Intelligent Tutorial-like” systems. In our proposed novel approach, *every* component incorporates the fundamental principles of LA. Thus, we model the Student (i.e., the learning mechanism) using an LA. Each Student is considered to be a member of a Classroom of Students, each of whom is individually represented by a distinct (and possibly different) LA. We also model the Domain and the Teacher using the LA paradigm.

Our research also works within a new philosophical perspective. We relax the constraint that “traditional” Tutorial systems have, namely the one of assuming that the Teacher is infallible. Rather, we assume that the Teacher is inherently uncertain of the domain knowledge, and is thus of a stochastic nature. However, although he is not absolutely certain about the material being taught, he is also capable of improving his *own* “teaching skills” even while the operation of the system proceeds. Finally, we also attempt to model a realistic learning framework, where the Students can learn not only from the Teacher, but also from other colleague Students in the Classroom.

Keywords: Tutorial-like Systems, Learning Automata, Modeling of Adaptive Systems.

1 Introduction

Central to every learning or adaptive system is an entity which performs the learning, and another entity which teaches the latter. Based on real-life analogies, these

¹ Since this a plenary talk, in the interest of readability, this document is written informally – with *minimal* mathematical formalism. More details of the mathematics, claims, potentials of the results, and the entire research endeavor can be found in the papers cited in the bibliography, and in the Doctoral Thesis of the second author [5].

can be informally referred to as the “Student” and the “Teacher” respectively. The aim of this research endeavor is to design a Tutorial-like system in which *every* component utilizes the fundamental principles of Learning Automata (LA). Indeed, we intend to model the Student (i.e., the learning mechanism) using an LA. We also propose to model a *Classroom* of Students - all of whom are appropriately represented, possibly by *different types* of LA. This, of course, broadens the horizons of both Tutorial systems and the fields of LA because we permit Students to not only learn from the Teacher, but to also learn by interacting with each other. Consequently, this research opens avenues for even more fascinating problems such as:

- How does the imperfect Teacher in such a Tutorial-like system present his² knowledge?
- How is the Domain knowledge represented?
- More importantly, since the Teacher himself is a component of the system, can we also incorporate learning capabilities in the Teacher, thus permitting him to improve his teaching capabilities as the the Student-Teacher interaction progresses?

This paper argues that LA-based Intelligent Tutorial-like systems can be an easy and useful way to model and implement Tutorial systems. We assume that the “Teacher” has an imprecise knowledge of the material to be imparted. While such a model for the Teacher has been studied extensively within the domains of LA, Neural Networks, and reinforcement learning [11,16,22], it is quite new to the field of Tutorial Systems. Thus, we believe that *after* our problem has been satisfactorily solved within a machine learning perspective, it can be, hopefully, ported to the application domain of Intelligent Tutorial systems.

The intention of the research in this paper is *not* to develop a generic system in which the Teacher is uncertain about the teaching material, or a generic model for the strategy by which he would impart the material and test the Students. Such a system would encounter enormous hurdles related to the psychological concepts of cognition, teaching, learning and intelligence, and also the system development aspects. In the research work presented in this paper, we propose that the problem we are studying be couched within the framework of the general machine learning paradigm, and thus we refer to the proposed system as a Tutorial-like system. Thus, it is reasonable for us to interchangeably refer to the “Student” as a “Student Simulator”, and to the “Teacher” as the “Teacher Simulator”, etc.

Using machine learning in improving tutoring systems was the study of a few previous researches. Frasson *et al.* [4] designed the main ITS components (student model, domain knowledge, and the tutoring model) in the form of intelligent agents. Lelouche [14] used a collection of interacting agents to represent the original modeling of the tutoring knowledge in an intelligent educational system. Legaspi and Sison [13] modeled the tutor in ITS using reinforcement

² For the ease of communication, we request the permission to refer to the entities involved (i.e. the Teacher, Student, etc.) in the masculine.

learning with the temporal difference method as the central learning procedure. Beck [3] used reinforcement learning to learn to associate superior teaching actions with certain states of the student's knowledge. Baffes and Mooney implemented ASSERT [2], which used reinforcement learning in student modeling to capture novel student errors using only correct domain knowledge. Our method is distinct from all the works cited here.

1.1 Tutorial-like Systems

Our entire research will be within the context of *Tutorial-like* systems [5]. In these systems, there need not be *real-life* Students, but rather each Student could be replaced by a Student Simulator that mimics a *real-life* Student. Alternatively, it could also be a software entity that attempts to learn. The Teacher, in these systems, attempts to present the teaching material to a *School* of Student Simulators. The Students are also permitted to share information between each other to gain knowledge. Therefore, such a teaching environment allows the Students to gain knowledge not only from the Teacher but also from other fellow Students.

In the *Tutorial-like* systems which we study, the Teacher has a *stochastic* nature, where he has an imprecise knowledge of the material to be imparted. The Teacher also doesn't have a prior knowledge about how to teach the subject material. He "learns" that himself while using the system, and thus, hopefully, improves his skills as a teacher. Observe that, conceptually, the Teacher, in some sense, is also a "student".

On the other hand, the Student Simulators need to learn from the Stochastic Teacher, as well as from each other. Each Student needs to decide when to request assistance from a fellow Student and how to "judge" the quality of information he receives from them. Thus, we require each Student to possess a mechanism whereby it can detect a scenario of procuring inaccurate information from other Students.

In our model of teaching/learning, the teaching material of the *Tutorial-like* system follows a Socratic model, where the domain knowledge is represented in the form of questions, either to be of a *Multiple Choice* sort or, in the most extreme case, of a *Boolean* sort. These questions, in our present paradigm, carry some degree of uncertainty, where each question has a probability that indicates the accuracy for the answer of that question.

1.2 Stochastic Learning Automaton

The stochastic automaton tries to reach a solution to a problem without any information about the optimal action. By interacting with an Environment, a stochastic automaton can be used to learn the optimal action offered by that Environment [11,12,15,16,17,20,24]. A random action is selected based on a probability vector, and then from the observation of the Environment's response, the action probabilities are updated, and the procedure is repeated. A stochastic automaton that behaves in this way to improve its performance is called a Learning Automaton (LA).

In the definition of a Variable Structure Stochastic Automata (VSSA), the LA is completely defined by a set of actions (one of which is the output of the automaton), a set of inputs (which is usually the response of the Environment) and a learning algorithm, T . The learning algorithm [16] operates on a vector (called *the Action Probability vector*)

$$P(t) = [p_1(t), \dots, p_r(t)]^T,$$

where $p_i(t)$ ($i = 1, \dots, r$) is the probability that the automaton will select the action α_i at time 't',

$$p_i(t) = \Pr[\alpha(t) = \alpha_i], i = 1, \dots, r, \text{ and it satisfies}$$

$$\sum_{i=1}^r p_i(t) = 1 \forall t.$$

Note that the algorithm $T : [0,1]^r \times A \times B \rightarrow [0,1]^r$ is an updating scheme where $A = \{\alpha_1, \alpha_2, \dots, \alpha_r\}$, $2 \leq r < \infty$, is the set of output actions of the automaton, and B is the set of responses from the Environment. Thus, the updating is such that

$$P(t+1) = T(P(t), \alpha(t), \beta(t)),$$

where $P(t)$ is the action probability vector, $\alpha(t)$ is the action chosen at time t , and $\beta(t)$ is the response it has obtained.

If the mapping T is chosen in such a manner that the Markov process has absorbing states, the algorithm is referred to as an absorbing algorithm [16] which are suitable for stationary environments. Ergodic VSSA better suited for non-stationary environments have also been investigated [16,19]. Furthermore, in order to increase their speed of convergence, the concept of discretizing the probability space was introduced [18,19]. This concept is implemented by restricting the probability of choosing an action to a finite number of values in the interval $[0,1]$.

Pursuit and Estimator-based LA were introduced to be faster schemes, since they pursue what can be reckoned to be the *current* optimal action or the set of current optimal schemes [19]. The updating algorithm improves its convergence results by using the history to maintain an estimate of the probability of each action being rewarded, in what is called the *reward-estimate* vector. Families of continuous and discretized Pursuit and Estimator-based LA have been shown to be faster than VSSA [23].

LA have been used in systems that have incomplete knowledge about the environment in which they operate. They have been used in telecommunications and telephone routing, image data compression, pattern recognition, graph partitioning, object partitioning, and vehicle path control [1].

1.3 Contributions of This Paper

In this research we propose a new philosophy to design and implement a Tutorial-like system in which *every* component utilizes the fundamental principles of LA. We present a novel design of a Tutorial-like system in which:

- The Teacher is stochastic, and uncertain about the teaching material.
- The Student is simulated to study the Domain knowledge.
- The Domain knowledge contains uncertain course material.

- The Teacher is dealing with a *School* of Students who learn from him and from each other.
- Since the Teacher himself is a specific component of the system, we also provide him with a mechanism to improve his own teaching abilities as system evolves.

In short, the goal of this research is to investigate how the field of Intelligent Tutorial systems and LA can be merged to produce Tutorial-*like* systems that have capabilities that are unreported in the literature.

2 Intelligent Tutorial/Tutorial-*like* Systems

An ITS consists of a number of modules, which are, typically, the domain model (knowledge domain), the student model, and the pedagogical model (which represent the tutor model itself). Self [21] defined these components as the tripartite architecture for an ITS – the *what* (domain model), the *who* (student model), and the *how* (tutoring model).

As mentioned in the introduction, Tutorial-*like* systems are quite similar (in principle) to traditional Tutorial systems, since they model the Teacher, the Student, and the Domain knowledge. However, there are many features which are distinct in the former, which is what we will briefly highlight below. More details of these distinctions can be found in [5].

1. **Different Type of Teacher.** As opposed to Tutorial systems, the Teacher in our Tutorial-*like* system possesses different features. First of all, one fundamental difference is that the Teacher is uncertain of the teaching material – he is stochastic. Also, the Teacher does not *initially* possess any knowledge about “How to teach” the domain subject. Rather, the Teacher himself is involved in a “learning” process and he “learns” what teaching material has to be presented to the Student.
2. **No Real Students.** A Tutorial system is intended for the use of *real-life* students. Its measure of accomplishment is based on comparing their progress with other students in a control group. Thus, the Tutorial-*like* system could be used by either:
 - (a) Students Simulators, that mimic the behavior and actions of *real-life* students using the system.
 - (b) An artificial Entity which, in itself, could be another software component that needs to “learn” specific domain knowledge.
3. **Uncertain Course Material.** Unlike the domain knowledge of “traditional” Tutorial systems where the knowledge is, typically, well defined, the domain knowledge for our Tutorial-*like* system contains material that has some degree of uncertainty.
4. **School of Students.** Traditional Tutorial Systems deal with a Teacher who teaches Students. A Tutorial-*like* system assumes that the Teacher is dealing with a *School* of Students where each learns from the Teacher on his own, and can also learn from his “colleagues” if he desires.

3 Overall Proposed Model

The overall proposed Tutorial-like system will be composed of an ensemble of LA modules, where each entity can improve its behavior based on the response it receives from *its corresponding* Environment. Thus, the system also uses LA to model the Student, the learning achieved by the Teacher, the learning achieved by each Student by interacting with the Teacher, and the learning accomplished by the School of Students by interacting between themselves. This paper presents an overall global view of our paradigm, and so we shall briefly (and without extensive mathematical formalism) describe each module in the Tutorial-like system.

To approach the general aim of this paper, we shall present the model for the overall Tutorial-like system, “piece by piece”, where each module uses stochastic LA. As discussed, our aim will be to use LA in every module of the system. However, the overall system is composed of different software modules which are capable of communicating with the model for the Teacher, and with each other. Figure 1 shows the different components of the system and their mutual interactions, described in greater detail, in the subsequent sub-sections.

3.1 The Model for the Student Simulator

In our Tutorial-like system, there is no *real-life* Student. Each Student is represented by a Student Simulator, which tries to mimic his behavior and actions. This is the actual entity that has to be taught the material by the Teacher. This modeling enables the Tutorial-like system to function and be tested without the need for *real-life* Students. The crucial features of the model involve the following facets:

- The Student Simulator is modeled using an LA paradigm. It uses LA as the learning mechanism to try to mimic the behavior and the learning of the Student.
- In addition to being able to learn from the Teacher, the Student Simulator is able to communicate with other Student Simulators, to enable it (them) to exchange information and learn from each other.

3.2 The Model for the Student

This model is a *representation* of the behavior and status of the Student (or the Student Simulator). It guides the Tutorial-like system to take tailored pedagogical decisions customized to each Student or his simulator. One can think of this as the model which the *system* retains in order to represent the Student, even though the Student may be learning using a completely different philosophy. Thus, for example, while the *real-life* Student may be learning using a neural network strategy, the system may model his learning paradigm using an L_{RI} LA scheme.

Attempting to understand how a learning mechanism within a “black box” learns is by no means trivial. Indeed, as far as we know, this is an unsolved problem. To achieve this, we shall use the philosophy briefly explained below.

- The model of the Student will be inferred using a higher-level of LA, referred to as the *Meta-LA*. While the Students use the Tutorial-like system, the

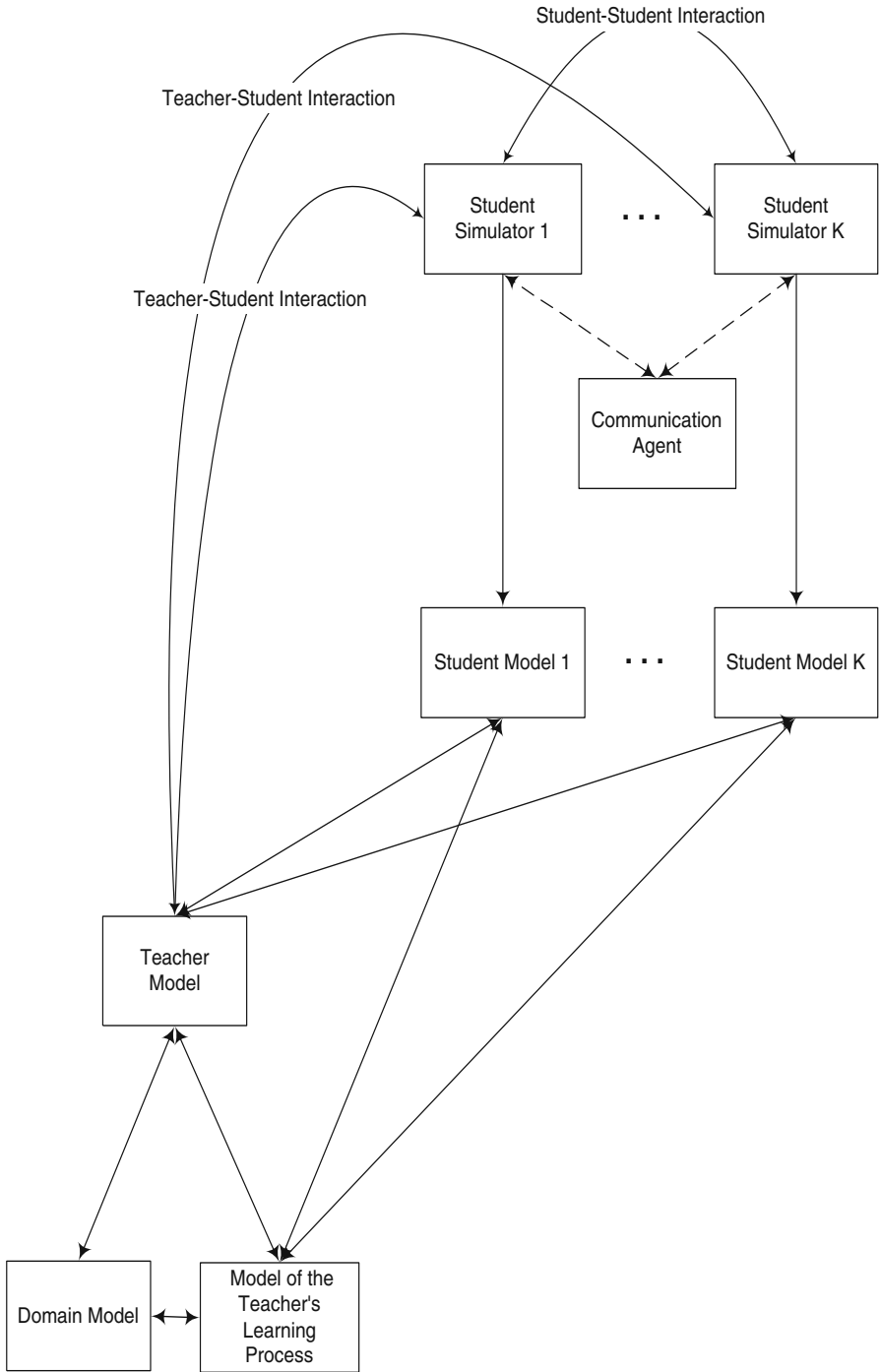


Fig. 1. Components of the LA-based Tutorial-like System

Meta-LA attempts to characterize their learning model. The input for the *Meta-LA* is obtained from its Environment, and does its task by observing the input/output characteristics of the Student Simulator.

- The *Meta-LA* will try to determine if the Student in question is one of the flowing types:
 - A Fast Learner.
 - Normal-paced Learner.
 - A Slow Learner.

Clearly, this sub-division of capabilities can be subject to a more fine resolution, but we believe that such a resolution is sufficient for most tutorial applications.

- The *Meta-LA* will determine the learning model of the Student by:
 - Observing and watching the consequent actions of the Student Simulator and his performance.
 - Examining the Environment’s characteristics, as well as observing its response to the Student’s actions.
- The *Meta-LA* Environment monitors the “performance” of the Student LA over a period of time. Depending on the sequence of observations, its Environment will penalize or reward the action of the *Meta-LA*.
- The higher-level *Meta-LA* and the lower-level LA are connected as a Network of LA that have a unique interaction model. In this model, the *Meta-LA* is not directly affected by the lower-level (Student) LA, but rather the *Meta-LA*’s Environment monitors the progress of the Student LA. Such a interconnection modeling is novel to the field of LA.

Although the description given here, for the model of the Student, is brief, additional details can be found in [5,6].

3.3 The Model for Student-Classroom Interaction

The Tutorial-like system allows the Student to be a member of a *Classroom* of students, where, for the most part, each member of the Classroom is permitted to not only learn from the Teacher(s) but also to “extract” information from any of his colleague students.

The Student-Classroom interaction is intended to maximize the learning experience of each Student as well as the collective learning of the Students. We propose a feasible model for the real-life scenario in which each Student can be provided with multiple sources of knowledge, each of which source is uncertain or inaccurate.

When interacting with each other, a Student, or the Student Simulator, can select an interaction strategy to communicate with other Students. Such a strategy can be one of the following:

1. He always assumes that the knowledge of other Students is reliable.

2. He assesses the knowledge received from other Students. If he considers this knowledge credible, he agrees to utilize this knowledge.
3. He initially accepts to give due consideration to the knowledge from other Students. He then evaluates the knowledge received after a period of probation, and “unlearns” the information gleaned if he infers that this information was misleading.
4. He decides to learn independently, and does not communicate with other fellow Students.

On the other hand, the Student defines how he is willing to take advantage of the knowledge that his colleagues can offer. Our model proposes two approaches in which the Student can handle this knowledge:

- A complete transfer of information, in which case the Student will acknowledge all the knowledge from the other Student, and forget or erase his learned state completely.
- A partial utilization of information, where the Student will use the knowledge from the other Student only to enhance his knowledge but not to erase the knowledge he has gained otherwise.

In order to facilitate the Student-Classroom interaction, the Student, or the Student Simulator, uses what we refer to as a *Tactic-LA*, to decide about the initiation of interaction steps. The *Tactic-LA* will enable the Student to consider one of the following methods of interaction:

- He is willing to offer his assistance to other Students in the Classroom.
- He is looking for assistance from other Students.
- He is not interested in any of the above options.

The Communication Agent is the component of the Tutorial-*like* system that is assigned to facilitate the interaction between Students. It enables the communication between Students by matching those Students who are willing to offer assistance, with their counterparts who request assistance.

In the context of the Student-Classroom interaction model, the results given in [5,8] can be summarized as follows: A Student-Classroom interaction is beneficial to most Students, especially the weaker ones, when they utilize the information they get from superior colleagues. Such a conclusion is, of course, intuitive, and we believe that our model is the pioneering one in this context.

More extensive details and experimental results for this can be found in [5,8].

3.4 The Model for the Domain

This model will encapsulate the Domain knowledge teaching material that needs to be communicated to the Students. It is the control center that encompasses the entire Domain knowledge, which the Teacher uses to impart knowledge to the Students. This module permits the Tutorial-*like* system to model and implement the Domain knowledge using a Socratic philosophy and *via* multiple-choice questions. It will represent the increasing complexity/difficulty of the material being taught. In essence, the features of this model will be as follows:

- The Domain knowledge will be presented *via* multiple-choice Socratic-type questions. For each question, every choice has an associated probability of being correct. The choice with the highest reward probability is the answer to that question.
- The knowledge imparted is arranged in *Chapters*, each of which will have a set of questions.
- Each *Chapter* represents a level of complexity/difficulty that is harder than the previous one.
- Students will not be able to predict the answer for subsequent *Chapters* using prior knowledge.

The Domain model we propose is capable of increasing the complexity of the Domain knowledge by reducing the range of the penalty probabilities of all actions (i.e., the multiple-choice answers), by incorporating a scaling factor μ . This will result in a clustering of the penalty probabilities, which will then lead to making the questions more complex inasmuch as it will be more difficult for the Student to infer the correct action.

The reader is requested to refer to [5,7] for more details on the model for the Domain.

3.5 The Model for the Teacher

In this module, we model how the Teacher can teach the Socratic-type Domain knowledge, *via* multiple-choice questions. This knowledge is also used to *test* the Students in the imparted knowledge. The aim of this part of the study is to illustrate how the Stochastic Teacher can not only present Socratic-type Domain knowledge to the Students, but also model the way by which he can assist them so as to be able to learn increasingly complex material. Modeling the Teacher involves the following concepts:

- The Teacher retrieves Domain knowledge from the Domain model, so as to present it to the Student.
- The Domain knowledge, as mentioned in Section 3.4, is of the Socratic-type and is stored in the Domain model, via multiple-choice questions.
- The Domain knowledge is modeled with increasing complexity, which is intended to represent the increasing complexity of subsequent *Chapters* that the Student encounters.
- The Students will learn the Domain knowledge from the questions presented to them, and from the feedback obtained by answering these questions.
- The Teacher possesses a formal strategy to improve the ability of the Students to learn more complex material. He does this by assisting them with additional information so as to handle the Domain knowledge as the difficulty of the latter increases.

In order for the Teacher to assist the Students to handle complex Domain knowledge, he provides them with *hints*. These hints will be imparted to the Students in the form of increasing the initial probability for one of the actions in the

Student Simulator’s action probability vector. The Teacher has the ability to control the probability that the Student correctly receives the *hint*.

The model for the Teacher is described in greater detail in [5.9].

3.6 The Model for Improving the Teacher’s Skills

Our Tutorial-*like* system allows the Teacher to “learn” and improve his “teaching skills” while he is using the system. This is accomplished by providing a higher-level LA, referred to as the *Meta-Teacher*, which will infer the required customization that the *Teacher* needs for the particular Student. The salient features of this model are as follows:

- The Teacher will utilize the Student model. As explained in Section 3.2, this will be done by using the *Meta-LA* to infer the learning model of the Student and his learning capabilities.
- The *Meta-Teacher*, as a higher-level learning concept, will try to infer the required customization that the *Teacher* needs for the specific Student.
- For each Environment that the Student is attempting to learn from, the Teacher will define his *own* standards for the specific Environment.
- The *Meta-Teacher* will make its inferences based on observing the progress of the Students at an intermediate stage of the learning.
- Based on the knowledge inferred from the *Meta-Teacher* and the Student model, the Teacher will be able to customize the teaching material presented, and provide the appropriate *hints* to each individual Student.
- This customization demonstrates the adaptability of the Teacher to the particular needs and skills of each Student.

More detailed mathematical and experimental information concerning the model for improving the Teacher’s skills are found in [5.10], and omitted here in the interest of brevity.

3.7 The Overall Prototype of the Tutorial-*like* System

In all brevity, the prototype attempts to incorporate all the features of the different models, described in the above sub-sections. The goal of the prototype is to provide a researcher with the software necessary to simulate the teaching/learning experience within the generalized framework proposed by the tenets of this research.

Typically, a researcher who uses the system, proceeds along the following steps to test the prototype:

- Specify the configuration parameters for the Students and the Classroom. These parameters serve to define the characteristics of the Student/Classroom who interact with the Teacher. This includes the identity of each Student, the type of Student that the Student Simulator is trying to mimic (either Fast, Normal, or Below-Normal), the interaction strategy used by the Student, the rate of learning for the Student, etc.

- Define the configuration of the Domain knowledge.
- Define the learning Environment, its characteristics and how its complexity increases with the different *Chapters*.
- Provide a mechanism by which the Teacher will provide *hints* to assist the Students.
- Define how the Teacher himself will improve his teaching skills and abilities.

At the end of the simulations, the researcher who uses the prototype will be able to graphically observe the progress of each Student. How each student has learned, and the effect of the interaction between the Students will be depicted in these graphs.

4 Conclusion

In this paper, we have pioneered a new class of systems, which merged the fields of Intelligent Tutorial Systems (ITS) and Learning Automata (LA). We refer to these systems as being “Intelligent Tutorial-like” systems. Our research has succeeded in developing such Tutorial-like systems where every module was composed using LA, and where *every* entity improved its behavior based on the response it received from *its* corresponding Environment. Although the details of the various modules is omitted in the interest of brevity, they are found in the Doctoral Thesis of the second author [5] and in a sequence of publications [6,8,7,9,10] that have concentrated on the individual modules themselves.

Our Tutorial-like system has incorporated different concepts that are novel and unique. In our system, every facet of the interaction involved a model (or a non *real-life* Student), and in which the design and implementation followed an established learning paradigm, where every module of the system utilized the fundamental principles of LA. First of all, the Student (i.e., the learning mechanism) was modeled using a LA. The *imperfect* Teacher was modeled using an LA Environment. The Classroom of Students was also modeled using LA, with each of them being represented by distinct, and possibly *different types* of LA.

Throughout the paper, we have argued that LA-based Intelligent Tutorial-like systems can serve to be an easy and useful way to model and implement system-based (as opposed to *real-life* Student-based) Tutorial systems. It is our belief that after our problem has been satisfactorily solved within a machine learning perspective, it can be, hopefully, ported to the application domain of Intelligent Tutorial systems.

The paper also presented a philosophic view of such Tutorial-like systems in which we relaxed the infallible constraint that the “Teacher is ‘perfect’”. By modeling the Teacher as a stochastic “Entity”, we were able to perceive it as an Environment or Medium in which the Student is learning.

We conclude this paper by stating that we do not claim to have solved all the problems associated with the new field of Tutorial-like systems. However, we do believe that we have taken a few small but significant and pioneering steps in the direction of laying the foundational groundwork for the future.

References

1. Agache, M., Oommen, B.J.: Generalized pursuit learning schemes: New families of continuous and discretized learning automata. *IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics* 32(6), 738–749 (2002)
2. Baffes, P., Mooney, R.: Refinement-based student modeling and automated bug library construction. *Journal of AI in Education* 7(1), 75–116 (1996)
3. Beck, J.: Learning to teach with a reinforcement learning agent. In: *Proceedings of The Fifteenth National Conference on AI/IAAI*, Madison, WI, p. 1185 (1998)
4. Frasson, C., Mengelle, T., Aïmeur, E., Gouardères, G.: An actor-based architecture for intelligent tutoring systems. In: Lesgold, A.M., Frasson, C., Gauthier, G. (eds.) *ITS 1996. LNCS*, vol. 1086, pp. 57–65. Springer, Heidelberg (1996)
5. Hashem, M.K.: *Learning Automata Based Intelligent Tutorial-like Systems*. PhD thesis, School of Computer Science, Carleton University, Ottawa, Canada (2007)
6. Hashem, M.K., Oommen, B.J.: On using learning automata to model a student's behavior in a tutorial-like system. In: Okuno, H.G., Ali, M. (eds.) *IEA/AIE 2007. LNCS (LNAI)*, vol. 4570, pp. 813–822. Springer, Heidelberg (2007)
7. Hashem, M.K., Oommen, B.J.: Using learning automata to model a domain in a tutorial-like system. In: *Proceedings of ICMLC 2007, the 2007 International Conference of Machine Learning and Cybernetics*, Hong Kong, August 2007, pp. 112–118 (2007)
8. Hashem, M.K., Oommen, B.J.: Using learning automata to model a student-classroom interaction in a tutorial-like system. In: *Proceedings of IEEE-SMC 2007, the 2007 IEEE International Conference on Systems, Man and Cybernetics*, Montreal, October 2007, pp. 1177–1182 (2007)
9. Hashem, M.K., Oommen, B.J.: Using learning automata to model the behavior of a teacher in a tutorial-like system. In: *Proceedings of IEEE-SMC 2007, the 2007 IEEE International Conference on Systems, Man and Cybernetics*, Montreal, October 2007, pp. 76–82 (2007)
10. Hashem, M.K., Oommen, B.J.: Using learning automata to model the “learning process” of the teacher in a tutorial-like system. In: *Proceedings of ISCIS 2007, the 2007 International Symposium on Computer and Information Sciences*, Paper No. 1.3.C-012, Ankara, Turkey (November 2007)
11. Haykin, S.: *Neural Networks: A Comprehensive Foundation*. IEEE Press/ Macmillan College Publishing Company, New York (1994)
12. Lakshmivarahan, S.: *Learning Algorithms Theory and Applications*. Springer, Heidelberg (1981)
13. Legaspi, R.S., Sison, R.C.: Modeling the tutor using reinforcement learning. In: *Proceedings of the Philippine Computer Science Congress (PCSC)*, pp. 194–196 (2000)
14. Lelouche, R.: A collection of pedagogical agents for intelligent educational systems. In: Gauthier, G., VanLehn, K., Frasson, C. (eds.) *ITS 2000. LNCS*, vol. 1839, pp. 143–152. Springer, Heidelberg (2000)
15. Najim, K., Poznyak, A.S.: *Learning Automata: Theory and Applications*. Pergamon Press, Oxford (1994)
16. Narendra, K.S., Thathachar, M.A.L.: *Learning Automata: An Introduction*. Prentice-Hall, New Jersey (1989)
17. Obaidat, M.S., Papadimitriou, G.I., Pomportsis, A.S.: Learning automata: Theory, paradigms, and applications. *IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics* 32(6), 706–709 (2002)

18. Oommen, B.J.: Absorbing and ergodic discretized two-action learning automata. *IEEE Transactions on Systems, Man, and Cybernetics SMC-16*, 282–293 (1986)
19. Oommen, B.J., Agache, M.: Continuous and discretized pursuit learning schemes: Various algorithms and their comparison. *IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics* 31, 277–287 (2001)
20. Poznyak, A.S., Najim, K.: *Learning Automata and Stochastic Optimization*. Springer, Berlin (1997)
21. Self, J.: The defining characteristics of intelligent tutoring systems research: ITSs care, precisely. *International Journal of AI in Education* 10, 350–364 (1999)
22. Sutton, R.S., Barto, A.G.: *Reinforcement Learning: An Introduction*. MIT Press, Cambridge (1998)
23. Thathachar, M.A.L., Sastry, P.S.: Varieties of learning automata: An overview. *IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics* 32(6), 711–722 (2002)
24. Thathachar, M.A.L., Sastry, P.S.: *Networks of Learning Automata: Techniques for Online Stochastic Optimization*. Kluwer Academic, Boston (2003)

Modeling and Simulating Empires: Toward a Game World Generator

Barry G. Silverman

Electrical and Systems Engineering Department,
University of Pennsylvania, Philadelphia, USA
`basil@seas.upenn.edu`

Abstract. This talk enumerates the challenges of building a game generator that works like the SimCity and/or empire building genre of games. This talk begins by describing a universally recurring socio-cultural “game” of inter-group competition for control of resources. It next describes efforts to author a game generator and software agents able to play the game as real humans would - which suggests the ability to study alternative ways to influence them, observe effects, and potentially understand how best to alter the outcomes of dysfunctional economies and potential conflict situations. I then examine two implemented game worlds (NonKin Village and FactionSim Countries). I conclude by arguing that substantial effort on game realism, best-of-breed social science models, and agent validation efforts is essential if analytic experiments are to effectively explore alternative ways to influence outcomes.

User-Centric and Intelligent Service Composition in Ubiquitous Computing Environments

In-Young Ko

Computer Science Department,
Korea Advanced Institute of Science and Technology (KAIST)
Daejeon, Korea
`iko@kaist.ac.kr`

Abstract. The advancement of service-oriented computing and mobile device technologies gives us new challenges to provide intelligent services in ubiquitous computing (ubicom) environments. User-centricity and dynamism support are the most essential requirements to meet those challenges. In this talk, I will introduce a user-centric and intelligent service composition framework that allows users to create their personalized ubicom applications that utilize service resources in a highly dynamic ubicom environments. The main features of our framework include: (1) task-oriented and spontaneous service composition; (2) dynamic service monitoring and reconfiguration; and (3) pervasive service retrieval and management. I will also explain our experiences of applying this framework to urban computing applications and intelligent service robots.

Author Index

- Abe, Hidenao II-24, II-252
Abe, Jair Minoro II-380, II-388
Abe, Jun II-664
Abu Bakar, Rohani Binti II-212
Adachi, Yoshinori II-689
Aguilera, Felipe II-480
Ahmadinia, Ali II-498, II-515
Akama, Seiki II-380
Alhashel, Ebrahim II-102, II-111
Allende, Héctor I-22
Allende-Cid, Héctor I-22
Anquetil, Nicolas II-364
Aoe, Junichi II-308
Aoki, Kumiko II-539
Aregita, Andoni I-269
Arslan, Tughrul II-515
- Baba, Norio II-761
Babenyshv, Sergey I-38, II-16
Badcock, Jeremy II-42
Bai, Yun I-70
Balachandran, Bala M. II-102, II-111
Baralis, Elena II-50
Belani, Hrvoje I-318
Bellas, Francisco I-122, I-351
Benatier, Vincent I-261
Bergamaschi, Sonia II-58
Bienvenido, José Fernando II-74
Borysewicz, Krzysztof II-135, II-151
Bouché, Philippe I-219
Bouchard, Carole I-293
Briand, Henri I-261
Brodka, Piotr II-455
Brunelli, Ricardo I-147
Buch, Norbert I-169
Bultey, Alexis I-87
- Cagliero, Luca II-50
Caiv, Yichuan II-447
Caliusco, María Laura II-66
Carvalho, Fábio Romeu de II-388
Cerquitelli, Tania II-50
Chaignaud, Nathalie I-309
Chang, Jaewoo I-130
- Chen, Duan-Yu II-439
Chen, Sin-Yu II-411
Chen, Yi II-447
Chiang, Chao-Hong II-411
Chiang, Shu-Yin II-721, II-729
Chiba, Saori II-580
Chmaj, Grzegorz I-103
Chowdhury, Nihad Karim I-130
Chyr, Wen-Li II-745
Coronel, Mauricio II-66
Cuzzocrea, Alfredo II-91
- d'Anjou, Alicia II-846
de Beuvron, François I-87
Dell, Robert F. I-326
Dey, Chris II-42
Ding, Haochen II-220
Ding, Yi II-705
do Prado, Hércules Antonio II-348,
II-364
Dufaux, Alain I-184
Dujovne, Luis E. I-301
Duro, Richard J. I-122, I-351
- Eastwood, Mark I-209
El-Bakry, Hazem M. I-333
El-Koursi, El-Miloudi I-184
- Faina, Andres I-351
Fakhfakh, Nizar I-184
Fernandez-Canque, Hernando II-498,
II-515
Ferneda, Edilson II-348, II-364
Freer, John II-498
Fujimura, Naomi II-623
Fujita, Osamu II-767
Fujita, Yoshikatsu II-278
Fukuda, Kyoko II-853
Fukuda, Takahisa II-599
Fukue, Yoshinori II-293
Fukumura, Yoshimi II-531, II-539
Fuwa, Yasushi II-555
- Gabrys, Bogdan I-209
Gao, Ya II-736

- Gareli, Matías II-66
 Gargouri, Faiez I-277
 Garza, Paolo II-50
 González B., Juan J. I-227
 Graña, Manuel I-95, II-846
 Guerra, Francesco II-58
 Guerrero, Luis A. II-480
 Guirado, Rafael II-74
- Hamada, Mohamed I-333
 Harada, Hiroshi II-607
 Harada, Kouji II-782
 Hasegawa, Mikio II-607
 Hasegawa, Naoki II-531
 Hashem, M. Khaled I-360
 Hashizume, Aoi II-648
 Hattori, Akira II-631
 Hayama, Tessai II-547
 Hayami, Haruo II-631
 Hayashi, Akihiro II-300
 Hayashi, Tatsuya II-789
 He, Lifeng II-705
 Hernandez, Carmen II-846
 Hien, Vu Thanh I-161
 Hintea, Sorin II-498, II-506
 Hirata, Kouichi II-490
 Hirokawa, Sachio II-464
 Ho, Tu Bao I-251
 Holze, Marc II-82
 Hori, Koichi II-664
 Horiike, Toshihiko II-472
 Hsieh, Jun-Wei II-411
 Hušek, Petr I-11
 Huang, Fay II-421
 Huang, Xu II-119
- Ichikawa, Kohei II-270
 Iino, Yurie II-464
 Imai, Hideyuki II-333, II-340
 Inada, Yutaka II-308
 Inuzuka, Nobuhiro II-672
 Ioroi, Shigenori II-631
 Isaac, Jacques I-122
 Ishida, Yoshiteru II-782, II-789,
 II-797, II-805
 Ishii, Naohiro II-639, II-681, II-697,
 II-713
 Ito, Hideaki II-591
 Ito, Kimihito II-490
 Ito, Nobuhiro II-639
- Itoh, Hidenori II-705
 Iwahori, Yuji II-689, II-705
 Iwashita, Motoi II-228
 Iwata, Kazunori II-639
- Jacot, Jacques I-184
 Jimbo, Takashi II-697
 Jimenez-Molina, Angel II-829
 Jinya, Koji II-767
 Jurasovic, Kresimir I-46
- Kaminaga, Hiroaki II-580
 Kanda, Tomomi II-713
 Kang, Byung-Seok II-829
 Kashiji, Shinkaku II-308
 Kasprzak, Andrzej I-103, I-112
 Kawaguchi, Masashi II-697
 Kawai, Yuji II-591
 Kawamura, Kousuke II-656
 Kawaoka, Tsukasa I-200
 Kazienko, Przemyslaw II-455
 Khoudour, Louahdi I-184
 Kim, Ikno II-159, II-166, II-174, II-181
 II-188
 Kim, Jun-Sung II-829
 Kino, Yasunobu II-300
 Kishimoto, Ariyuki II-244
 Kitamura, Kentarou II-797
 Kitamura, Satoshi II-523
 Klawonn, Frank I-235
 Ko, In-Young I-375, II-829
 Kojima, Masanori II-773
 Kojiri, Tomoko II-564
 Koo, Hyung-Min II-829
 Koshimizu, Hiroyasu II-591
 Koszalka, Leszek I-112
 Kotowicz, Jean-Philippe I-309
 Kołaczek, Grzegorz I-30
 Krömer, Pavel I-139
 Kuboyama, Tetsuji II-472, II-490
 Kudo, Mineichi II-333
 Kudo, Yasuo II-356
 Kunifuji, Susumu II-547
 Kunimune, Hisayoshi II-555
 Kusek, Mario I-46
 Kwasnicka, Halina II-135, II-151
- Le, Bac I-251
 Lee, Hsuan-Shih II-745
 Lee, Huey-Ming II-195

- Lee, Hyunjo I-130
 Lehn, Remi I-261
 Lenzen, Manfred II-42
 Liao, Hong-Yuan Mark II-421
 Lin, Daw-Tung II-431
 Lin, Kuang II-745
 Lin, Lily II-195
 Liu, Li-Wei II-431
 Lopez, Ruben I-122
 Lortal, Gaëlle I-309
 Lovrek, Ignac I-46
 Lu, Jie II-372, II-736
 Lu, Ning II-372
- Magalhães, André Ribeiro II-348
 Majdoubi, Jihen I-277
 Makino, Toshiyuki II-672
 Makris, Dimitrios I-169
 Marchetti, Marco II-50
 Markowska-Kaczmar, Urszula II-151,
 II-861
 Martín, Juanxu I-269
 Martínez F., José A. I-227
 Masui, Takahiro II-572
 Matumura, Tetuya II-773
 Meghdadi, A.H. II-127
 Mendoza, Marcelo I-285
 Mineno, Hiroshi II-572, II-648
 Miranda, Clara Marcela II-74
 Mitani, Keiichiro II-293
 Mitsuishi, Takashi II-236
 Miura, Hajime II-531
 Miura, Motoki II-547
 Miyadera, Youzou II-580
 Miyahara, Shiori II-523
 Miyata, Masako II-664
 Miyazaki, Toshimasa II-531
 Mizuno, Shinji II-539
 Mizuno, Tadanori II-572, II-648, II-773
 Mizutani, Masayuki II-681
 Mohammedian, Masoud II-102
 Monroy, Juan I-122
 Moradian, Esmiralda II-821
 Moraga, Claudio I-22
 Morales-Rodríguez, María Lucila I-227
 Moreno, Ramón II-846
 Morita, Hiroyuki II-259
 Mukai, Naoto II-656
 Murai, Tetsuya II-356
- Musa, Zalili Binti II-220
 Musial, Katarzyna II-455
- Nakachi, Namiko II-270
 Nakahara, Takanobu II-259
 Nakamatsu, Kazumi II-380, II-388
 Nakamura, Shoichi II-580
 Nakamura, Tsuyoshi II-705
 Nakano, Hideo II-308
 Nara, Yumiko II-32
 Nath, Rudra Pratap Deb I-130
 Nguyen, Huy I-251
 Nguyen, Ngoc Thanh I-54
 Nicolicin-Georgescu, Vlad I-261
 Nin, Kou II-761
 Nishihara, Yoko II-1
 Nishimatsu, Ken II-228
 Nishino, Kazunori II-539
 Nucita, Andrea II-91
 Numa, Kosuke II-664
- Obata, Kenji II-572
 Ogiela, Lidia I-177
 Ogiela, Marek R. I-177, I-192
 Ohia, Dawid I-112
 Ohsawa, Yukio II-1, II-9
 Okada, Yousuke II-681
 Oltean, Gabriel II-506
 Oommen, B. John I-360
 Orsini, Mirko II-58
 Orwell, James I-169
 Ozaki, Masahiro II-689
- Paradowski, Mariusz II-135, II-151
 Pavón, Ruth I-147
 Pavešić, Nikola I-1
 Pazos R., Rodolfo A. I-227
 Pécuchet, Jean-Pierre I-309
 Peters, J.F. II-127
 Pham, Tuan D. I-155
 Platoš, Jan I-139
 Posada, Jorge I-95
 Pozniak-Koszalka, Iwona I-103
 Prieto, Abraham I-351
 Pripuzić, Krešimir I-318
 Puentes, Antonio I-122
- Ramanna, S. II-127
 Rebolledo, L. Víctor II-838
 Ren, Fuji I-62
 Ribarić, Slobodan I-1

- Ríos, Sebastián A. II-480
 Ritter, Norbert II-82
 Rodriguez, Marko A. II-813
 Rojas P., Jessica C. I-227
 Román, Pablo E. I-326
 Rousselot, François I-87
 Rubiolo, Mariano II-66
 Rybakov, Vladimir I-38
 Rybakov, Vladimir V. II-16
- Sakamoto, Hiroshi II-472
 Sakurai, Hirohito II-340
 Salas, Rodrigo I-22
 Sanín, Cesar I-95
 Sartori, Claudio II-58
 Saruwatari, Yasufumi II-278
 Sata, Kazuya II-490
 Sato, Yoshiharu II-340
 Setchi, Rossitza I-293
 Setoguchi, Yoichi II-639
 Sharma, Dharmendra II-111, II-119
 Shen, Pei-Di II-745
 Shidama, Yasunari II-236
 Shih, Timothy K. II-421
 Shimogawa, Shinsuke II-228
 Shimohara, Katsunori II-752
 Shin, Kilho I-342
 Shirai, Hirokazu II-580
 Silverman, Barry G. I-374
 Sipos, Emilia II-506
 Śluzek, Andrzej II-143
 Snášel, Václav I-139
 Soga, Masato II-599
 Stegmayer, Georgina II-66
 Straccia, Umberto I-78
 Su, Chih-Wen II-421
 Sugimoto, Tatsuo II-664
 Suzuki, Nobuo II-317
 Suzuki, Shoji II-697
 Switek, Tomasz II-861
 Szczerbicki, Edward I-95
- Tadeusiewicz, Ryszard I-177, I-192
 Takahashi, Hiroataka II-531
 Takahashi, Masakazu II-244, II-278,
 II-285
 Takahashi, Toru II-244
 Takahashi, Youhei II-472
 Takahashi, Yuichi II-1
 Takamiya, Masatoshi II-615
- Takeda, Taichi II-607
 Takeuchi, Akiko II-531
 Taki, Hirokazu II-599
 Takizawa, Takeshi II-555
 Tanaka, Takushi II-325
 Tanev, Ivan II-752
 Tang, Nick C. II-421
 Tang, Qiao I-293
 Teixeira, Elizabeth d'Arrochella II-364
 Terano, Takao II-244, II-285
 Tmar, Mohamed I-277
 Tokuda, Yusaku II-713
 Tokumitsu, Masahiro II-797
 Tomobe, Hironori II-664
 Torii, Ippei II-681, II-713
 Toriumi, Kiyoko II-664
 Toro, Carlos I-95, I-269
 Toya, Hiroko II-539
 Toyama, Jun II-333
 Tran, Trong Hieu I-54
 Trzupek, Mirosław I-192
 Tsai, Wen-Hsiang II-395, II-403
 Tschumitschew, Katharina I-235
 Tsuchiya, Seiji I-62, I-200, I-243
 Tsuda, Kazuhiko II-278, II-285, II-293,
 II-300, II-317
 Tsumoto, Shusaku II-24, II-252
- Ubukata, Seiki II-356
 Ushiyama, Taketoshi II-623
- Vaquero, Javier I-95, I-269
 Velásquez, Juan D. I-301, I-326, II-838
 Velastin, Sergio A. I-169
 Veloz, Alejandro I-22
 Vincini, Maurizio II-58
 Vo, Bay I-251
 von Lücken, Christian I-147
 Vuković, Marin I-318
- Wang, Jin-Long II-721, II-729
 Washio, Takashi II-270
 Watabe, Hirokazu I-62, I-200, I-243
 Watada, Junzo II-159, II-166, II-174,
 II-181, II-188, II-203,
 II-212, II-220
 Watanabe, Toyohide II-564, II-615
 Watanabe, Yuki II-564
 Watari, Shinichi II-797
 Watkins, Jennifer H. II-813

- Woodham, Robert J. II-705
Wu, Chih-Jen II-395, II-403
Xu, Xian Ming I-342
Yaakob, Shamshul Bahar II-203
Yada, Katsutoshi II-270
Yamada, Koji II-547
Yamada, Kozo II-340
Yamada, Kunihiro II-773
Yamada, Masanori II-523
Yamada, Takashi II-244
Yamamoto, Kosuke II-615
Yamauchi, Yuhei II-523
Yamazaki, Makoto II-531
Yan, Sun II-220
Yano, Shohei II-531
Yatsugi, Kotaro II-623
Yin, Fei I-169
Yokoyama, Kenzou II-555
Yokoyama, Setsuo II-580
Yoshida, Kouji II-773
Yoshimura, Eriko I-62, I-200, I-243
Yu, Ting II-42
Yukawa, Takashi II-531
Zadrija, Valentina I-1
Zamora, Juan I-285
Zanni-Merk, Cecilia I-87, I-219
Zhang, Guangquan II-372, II-736
Zhang, Guoli II-736