LNAI 5178

Ignac Lovrek
Robert J. Howlett
Lakhmi C. Jain (Eds.)

# Knowledge-Based Intelligent Information and Engineering Systems

12th International Conference, KES 2008
Zagreb, Croatia, September 2008
Proceedings, Part II

2 Part II

Springer

Lecture Notes in Artificial Intelligence 5178

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

Subseries of Lecture Notes in Computer Science

Ignac Lovrek Robert J. Howlett
Lakhmi C. Jain (Eds.)

# Knowledge-Based Intelligent Information and Engineering Systems

12th International Conference, KES 2008
Zagreb, Croatia, September 3-5, 2008
Proceedings, Part II

Springer

Volume Editors

Ignac Lovrek
University of Zagreb, Faculty of Electrical Engineering and Computing
Unska 3, 10000 Zagreb, Croatia
E-mail: ignac.lovrek@fer.hr

Robert J. Howlett
KES International
2nd Floor, 145-157 St. John Street, London EC1V 4PY, UK
E-mail: rjhowlett@kesinternational.org

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

Library of Congress Control Number: 2008933454

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 3-540-85564-5 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-85564-4 Springer Berlin Heidelberg New York

Springer is a part of Springer Science+Business Media

springer.com

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 12511358 06/3180 5 4 3 2 1 0

# Preface

Delegates and friends, we are very pleased to extend to you a warm welcome to this, the 12th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems organised by the Faculty of Electrical Engineering and Computing at the University of Zagreb, in association with KES International.

For over a decade, KES International has provided an annual wide-spectrum intelligent systems conference for the applied artificial intelligence research community. Having originated in Australia and been held there during 1997–99, the conference visited the UK in 2000, Japan in 2001, Italy in 2002, the UK in 2003, New Zealand in 2004, Australia in 2005, the UK in 2006, Italy in 2007, and now in Zagreb, Croatia in 2008. It is planned that KES 2009 will be held in Santiago, Chile before returning to the UK in 2010. The KES conference is mature and regularly attracts several hundred delegates. As it encompasses a broad range of intelligent systems topics, it provides delegates with an opportunity to mix with researchers from other groups and learn from them. The conference is linked to the International Journal of Intelligent and Knowledge-Based Systems, published by IOS Press under KES editorship. Extended and enhanced versions of the best papers presented at the KES conference may be published in the Journal.

In addition to the annual wide-range intelligent systems conference, KES has run successful symposia in several specific areas of the discipline. Agents and Multi-Agent Systems is a popular area of research. The first KES symposium on Agents and Multi-Agent Systems took place in 2007 in Wroclaw, Poland (KES-AMSTA 2007) followed in 2008 by a second in Incheon, Korea (KES-AMSTA 2008). The third in the series is planned to be held in the historic city of Uppsala in Sweden (June 3–5, 2009). Intelligent Multi-Media is a second area of focus for KES. The first KES symposium on Intelligent and Interactive Multi-Media Systems and Services (KES IIMSS 2008) will be held in Athens, Greece, in 2008, followed by a second in Venice, Italy, in 2009 (dates to be notified). A third area of interest supported by KES is Intelligent Decision Support Technologies, and the first KES symposium on this subject (KES IDT 2009) is planned for Hyogo in Japan for next year (April 23–25, 2009). Over time, each of these areas will be supported by a KES focus group of researchers interested in the topic, and if appropriate, by a journal. To this end, the International Journal of Intelligent Decision Support Technologies is published by IOS Press under the editorship of a developing KES IDT focus group.

For the future we have plans and a vision for KES. Firstly, we describe the plans.

We plan to maintain and increase the quality of KES publications. The KES quality principle is that we do not seek to expand KES activities by publishing inferior papers. However, equally, we do not believe it serves authors or the research community to reject good papers on the basis of an arbitrary acceptance / rejection ratio. Hence papers for KES conferences and symposia, and the KES Journal, will be rigorously reviewed by experts in the field, and published only if they are of a sufficiently high level, judged by international research standards.

We will further develop KES focus groups, and where appropriate, we will adopt journals and symposia, where this supports and helps us maintain our quality principle.

We introduced the concept of KES membership several years ago to provide returning KES delegates with discounted conference fees. We plan to supplement the benefits of KES membership by launching a profile page system, such that every KES member will have their own profile page on the KES web site, and be able to upload a description of themselves, their research interests and activities. The member site will act as a contact point for KES members with common interests and a potential channel to companies interested in members' research.

Printing technology is changing and this will have an effect on publishers and publications. We are conducting trials with rapid publication technology that makes it possible to print individual copies of a book on demand. The KES Rapid Research Results book series will make it convenient to publish books appealing to niche markets, for example specialised areas of research, in a way that would not have been economic before.

Many KES members and supporters have research interests outside intelligent systems. In fact, intelligent systems may just be a tool used for an application which is the main interest. A significant number of those involved in KES have interests in environmental matters. In 2009, KES will address the issue of sustainability and renewable energy through its first conference in this area, Sustainability in Energy and Buildings (SEB 2009), which will be held in Brighton, UK, April 30–May 1, 2009. SEB 2009 will address a broad spectrum of sustainability issues relating to renewable energy and the efficient use of energy in domestic and commercial buildings. Papers on the application of intelligent systems to sustainability issues will be welcome. However, it will not be compulsory that papers for SEB 2009 have significant intelligent systems content (as is a criterion for other KES conferences and symposia).

In addition to the firm plans for KES there is a longer term vision. In the time that it has been in existence, KES International has evolved from being the organiser of just a single annual conference, to a provider of an expanding portfolio of support functions for the research community. Undoubtedly, KES will continue to develop and enhance its knowledge transfer activities. The KES community consists of several thousand members, and potentially it could play a significant role in generating synergy and facilitating international research co-operation. A long term vision for KES is for it to evolve into an international academy providing the means for its members to perform international collaborative research projects. At the moment we do not have the means to turn this vision into a reality, but we will work towards this aim.

The annual KES conference continues to be a major feature of the KES organisation, and KES 2008 will continue the tradition of excellence in KES conferences.

The papers for KES 2008 were submitted either to Invited Sessions, chaired and organised by respected experts in their fields, or to General Sessions, managed by Track Chairs. Each paper was thoroughly reviewed by two members of the International Review Committee, and also inspected by a Track Chair or Invited Session Chair. A decision about whether to publish the paper was made, based on the KES quality principle described above. If the paper was judged to be of high enough quality to be accepted, the Programme Committee then decided on oral or poster presentation, based on the subject and content of the paper. All papers at KES 2008 are considered to be of equal weight and importance, no matter whether they were oral or poster presentations. This has resulted in the 316 high-quality papers included in these proceedings.

Thanks are due to the very many people who have given their time and goodwill freely to make the conference a success.

We would like to thank the KES 2008 International Programme Committee for their help and advice, and also the International Review Committee, 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 expert talks to inform and inspire subsequent discussions.

An important feature of KES conferences is the Invited Session Programme. Invited Sessions give both young and established researchers an opportunity to organise and chair a set of papers on a specific topic, presented as a themed session. In this way, new topics at the leading edge of intelligent systems can be presented to interested delegates. This mechanism for feeding new ideas into the research community is very valuable. We thank the Invited Session Chairs who have contributed in this way.

The conference administrators, and the local organising committee, have all worked extremely hard to bring the conference to a high level of organisation. In this context, we would like to thank Mario Kusek, Kresimir Jurasovic, Igor Ljubi, Ana Petric, Vedran Podobnik and Jasna Slavinic (University of Zagreb, Croatia); Peter Cushion, Nicola Pinkney and Antony Wood (KES Operations, UK).

A vital contribution was made by the authors, presenters and delegates without whom the conference could not have taken place. So finally, but by no means least, we thank them for their involvement.

June 2008

Bob Howlett
Ignac Lovrek
Lakhmi Jain

# Organisation

## KES 2008 Conference Organisation

KES 2008 was organised by KES International, Innovation in Knowledge-Based and Intelligent Engineering Systems, and the University of Zagreb, Faculty of Electrical Engineering and Computing.

## KES 2008 Conference Chairs

| | |
|---|---|
| General Chair | Ignac Lovrek, University of Zagreb, Croatia |
| Executive Chair | Robert J. Howlett, University of Brighton, UK |
| Invited Sessions Chair | Lakhmi C. Jain, University of South Australia, Australia |
| Local Chair | Mario Kusek, University of Zagreb, Croatia |
| Award Chair | Bogdan Gabrys, University of Bournemouth, UK |

## KES Conference Series

KES 2008 is a part of the KES Conference Series.

| | |
|---|---|
| Conference Series Chairs | Lakhmi C. Jain and Robert J. Howlett |
| KES Executive Chair | Robert J. Howlett, University of Brighton, UK |
| KES Founder | Lakhmi C. Jain, University of South Australia, Australia |

## Local Organising Committee

Kresimir Jurasovic, Igor Ljubi, Ana Petric, Vedran Podobnik, Jasna Slavinic (University of Zagreb, Croatia), Peter Cushion (KES Operations, UK)

## International Programme Committee

| | |
|---|---|
| Abe, Akinori | ATR Knowledge Science Laboratories, Japan |
| Adachi, Yoshinori | Chubu University, Japan |
| Angulo, Cecilio | Technical University of Catalonia, Spain |
| Apolloni, Bruno | University of Milan, Italy |
| Baba, Norio | Osaka Kyoiku University, Japan |
| Balachandran, Bala M. | University of Canberra, Australia |
| Dalbelo Basic, Bojana | University of Zagreb, Croatia |
| Beristain, Andoni | Universidad del Pais Vasco, Spain |

| | |
|---|---|
| Bianchini, Monica | Universita degli Studi di Siena, Italy |
| Castellano, Giovanna | University of Bari, Italy |
| Chen, Yen-Wei | Ritsumeikan University, Japan |
| Cheng, Jingde | Saitama University, Japan |
| Corchado, Emilio | University of Burgos, Spain |
| Cuzzocrea, Alfredo | University of Calabria, Italy |
| Damiani, Ernesto | University of Milan, Italy |
| Di Noia, Tommaso | Technical University of Bari, Italy |
| Esposito, Floriana | University of Bari, Italy |
| Gabrys, Bogdan | University of Bournemouth, UK |
| Gao, Kun | Zhejiang Wanli University, China |
| Hartung, Ronald L. | Franklyn University, USA |
| Hakansson, Anne | Uppsala University, Sweden |
| Holmes, Dawn | University of California Santa Barbara, USA |
| Howlett, Robert J. | University of Brighton, UK |
| Ishida, Yoshiteru | Toyohashi University of Technology, Japan |
| Ishii, Naohiro | Aichi Institute of Technology, Japan |
| Jain, Lakhmi C. | University of South Australia, Australia |
| Jevtic, Dragan | University of Zagreb, Croatia |
| Karny, Miroslav | Academy of Sciences of the Czech Republic, Czech Republic |
| Kunifuji, Susumu | Japan Advanced Institute of Science and Technology, Japan |
| Lee, Hsuan-Shih | National Taiwan Ocean University, Taiwan |
| Lim, Chee-Peng | University of Science, Malaysia |
| Liu, Honghai | University of Portsmouth, UK |
| Lovrek, Ignac | University of Zagreb, Croatia |
| Maojo, Victor | Universidad Politecnica de Madrid, Spain |
| Markey, Mia | University of Texas at Austin, USA |
| Mumford, Christine | Cardiff University, UK |
| Munemori, Jun | Wakayama University, Japan |
| Nakamatsu, Kazumi | University of Hyogo, Japan |
| Nakano, Ryohei | Nagoya Institute of Technology, Japan |
| Nakao, Zensho | University of the Ryukyus, Japan |
| Nauck, Detlef | BT Intelligent Systems Research Centre, UK |
| Negoita, Mircea Gh. | KES International |
| Nguyen, Ngoc Thanh | Wroclaw University of Technology, Poland |
| Nicoletti, Maria do Carmo | Federal University of Sao Carlos, Brazil |
| Nikolos, Ioannis K. | Technical University of Crete, Greece |
| Nishida, Toyoaki | Kyoto University, Japan |
| Nuernberger, Andreas | University of Magdeburg, Germany |
| Palade, Vasile | University of Oxford, UK |
| Park, Gwi-Tae | Korea University, Seoul, Korea |
| Pham, Tuan | James Cook University, Australia |
| Phillips-Wren, Gloria | Loyola College in Maryland, USA |

| | |
|---|---|
| Sharma, Dharmendra | University of Canberra, Australia |
| Shkodirev, Viacheslaw | St Petersburg State Polytechnic University, Russia |
| Sik Lanyi, Cecilia | University of Pannonia, Hungary |
| Sunde, Jadranka | Defence Science and Technology Organisation, Australia |
| Tagliaferri, Roberto | University of Salerno, Italy |
| Taki, Hirokazu | Wakayama University, Japan |
| Tsuda, Kazuhiko | University of Tsukuba, Japan |
| Turchetti, Claudio | Università Politecnica delle Marche, Italy |
| Vassanyi, Istvan | University of Pannonia, Hungary |
| Veganzones, Miguel A. | Universidad del Pais Vasco, Spain |
| Vellido, Alfredo | Technical University of Catalonia, Spain |
| Watada, Junzo | Waseda University, Japan |
| Watanabe, Toyohide | Nagoya University, Japan |
| Yamashita, Yoshiyuki | Tokio University of Agriculture and Technology, Japan |

## International Review Committee

| | |
|---|---|
| Abe, Akinori | ATR Knowledge Science Laboratories, Japan |
| Abe, Jair | University of Sao Paulo, Brazil |
| Abu Bakar, Rohani | Waseda University, Japan |
| Abulaish, Muhammad | Jamia Millia Islamia, India |
| Adachi, Yoshinori | Chubu University, Japan |
| Adli, Alexander | University of the Ryukyus, Japan |
| Akama, Seiki | C-Republics, Japan |
| Al-Hashel, Ebrahim | University of Canberra, Australia |
| Alquezar, Ren | Technical University of Catalonia, Spain |
| Angelov, Plamen | Lancaster University, UK |
| Anguita, Davide | University of Genoa, Italy |
| Angulo, Cecilio | Technical University of Catalonia, Spain |
| Anisetti, Marco | University of Milan, Italy |
| Aoyama, Kouji | Fujitsu Laboratories Limited, Japan |
| Apolloni, Bruno | University of Milan, Italy |
| Appice, Annalisa | University of Bari, Italy |
| Aritsugi, Masayoshi | Kumamoto University, Japan |
| Arroyo-Figueroa, Gustavo | Instituto de Investigaciones Electricas, Mexico |
| Azzini, Antonia | University of Milan, Italy |
| Baba, Norio | Osaka Kyoiku University, Japan |
| Balachandran, Bala M. | University of Canberra, Australia |
| Balas, Marius | Aurel Vlaicu University of Arad, Romania |
| Balas, Valentina E. | Aurel Vlaicu University of Arad, Romania |
| Balic, Joze | University of Maribor, Slovenia |
| Bandini, Stefania | University of Milan Bicocca, Italy |
| Baruque, Bruno | University of Burgos, Spain |

| | |
|---|---|
| Basili, Roberto | University of Rome, Italy |
| Bassis, Simone | University of Milan, Italy |
| Bayarri, Vicente | GIM Geomatics S.L., Spain |
| Belanche, Lluis | Technical University of Catalonia, Spain |
| Bellandi, Valerio | University of Milan, Italy |
| Berendt, Bettina | Katholieke Universiteit Leuven, Belgium |
| Bianchini, Monica | Università degli Studi di Siena, Italy |
| Bidlo, Michal | Technical University of Brno, Czech Republic |
| Bielikov, Mria | Slovak University of Technology, Slovakia |
| Billhardt, Holger | Universidad Rey Juan Carlos I, Spain |
| Bingul, Zafer | Kocaeli University, Turkey |
| Bioucas, Jose | Instituto Superior Tecnico, Portugal |
| Bogdan, Stjepan | University of Zagreb, Croatia |
| Borghese, Alberto | University of Milan, Italy |
| Borzemski, Leszek | Wroclaw University of Technology, Poland |
| Bouchachia, Abdelhamid | University of Klagenfurt, Austria |
| Bouquet, Paolo | Università degli Studi di Trento, Italy |
| Bouridane, Ahmed | Queen's University, Belfast, UK |
| Bruzzone, Lorenzo | Università degli Studi di Trento, Italy |
| Buciu, Ioan | University of Oradea, Romania |
| Cabestany, Joan | Technical University of Catalonia, Spain |
| Calpe-Maravilla, Javier | University of Valencia, Spain |
| Camino-Gonzalez, Carlos Luis | Forestry and Technology Center of Catalonia, Spain |
| Camps-Valls, Gustavo | University of Valencia, Spain |
| Cao, Jiangtao | University of Portsmouth, UK |
| Capkovic, Frantisek | Slovak Academy of Sciences, Slovakia |
| Carpintero-Salvo, Irene Rosa | Universidad de Granada and Consejeria de Medio Ambiente, Spain |
| Castellano, Giovanna | University of Bari, Italy |
| Castiello, Ciro | University of Bari, Italy |
| Castillo, Elena | University of Cantabria, Spain |
| Cavar, Damir | Indiana University, USA |
| Ceccarelli, Michele | University of Sannio, Italy |
| Ceravolo, Paolo | University of Milan, Italy |
| Chan, Chee Seng | University of Portsmouth, UK |
| Chang, Chan-Chih | Industrial Technology Research Institute, Taiwan |
| Chang, Chuan-Yu | National Yunlin University of Science & Technology, Taiwan |
| Chen, Mu-Yen | National Changhua University of Education, Taiwan |
| Chen, Yen-Wei | Ritsumeikan University, Japan |
| Cheng, Jingde | Saitama University, Japan |
| Chetty, Girija | University of Canberra, Australia |

| | |
|---|---|
| Chi Thanh, Hoang | Hanoi University of Science, Vietnam |
| Ciaramella, Angelo | University of Naples Parthenope, Italy |
| Cicin Sain, Marina | University of Rijeka, Croatia |
| Claveau, Vincent | IRISA-CNRS, France |
| Coghill, George | University of Auckland, New Zealand |
| Colucci, Simona | Technical University of Bari, Italy |
| Corbella, Ignasi | Technical University of Catalonia, Spain |
| Corchado Rodriguez, Juan M. | University of Salamanca, Spain |
| Corchado, Emilio | University of Burgos, Spain |
| Costin, Mihaela | Institute for Computer Science, Romanian Academy, Romania |
| Cox, Robert | University of Canberra, Australia |
| Crippa, Paolo | Università Politecnica delle Marche, Italy |
| Cruz, Antonio | Federal University of Sao Carlos, Brazil |
| Csipkes, Gabor | Technical University of Cluj-Napoca, Romania |
| Curzi, Alessandro | Università Politecnica delle Marche, Italy |
| Cuzzocrea, Alfredo | University of Calabria, Italy |
| Czarnowski, Ireneusz | Gdynia Maritime University, Poland |
| D'Amato, Claudia | University of Bari, Italy |
| D'Apuzzo, Livia | University of Naples Federico II, Italy |
| Dalbelo Basic, Bojana | University of Zagreb, Croatia |
| Damiani, Ernesto | University of Milan, Italy |
| Davidsson, Paul | Blekinge Institute of Technology, Sweden |
| de Campos, Cassio Polpo | Rensselaer Polytechnic Institute, USA |
| De Gemmis, Marco | University of Bari, Italy |
| De Santis, Angela | University of Alcala, Spain |
| Deguchi, Toshinori | Gifu National College of Technology, Japan |
| Dell'Endice, Francesco | University of Zurich, Switzerland |
| Di Noia, Tommaso | Technical University of Bari, Italy |
| Di Sciascio, Eugenio | Technical University of Bari, Italy |
| Diani, Marco | University of Pisa, Italy |
| Diaz-Delgado, Ricardo | Estacion Biologica de Donana-CSIC, Spain |
| Dobsa, Jasminka | University of Zagreb, Croatia |
| Dorado, Julian | Universidad de la Coruna, Spain |
| Dujmic, Hrvoje | University of Split, Croatia |
| Dumitriu, Luminita | Dunarea de Jos University, Romania |
| Duro, Richard | Universidade da Coruna, Spain |
| Edman, Anneli | Uppsala University, Sweden |
| Erjavec, Tomaz | Josef Stefan Institute, Slovenia |
| Esposito, Anna | University of Naples Federico II and IIASS, Italy |
| Esposito, Floriana | University of Bari, Italy |
| Etchells, Terence | Liverpool John Moores University, UK |
| Lee, Eun-Ser | Soong Sil University, Korea |
| Fang, H.H. | Taipei College of Maritime Technology, Taiwan |

| | |
|---|---|
| Fasano, Giovanni | University of Venice, Italy |
| Feng, Jun | Hohai University, China |
| Fernandez-Caballero, Antonio | Universidad de Castilla-La Mancha, Spain |
| Fras, Mariusz | Wroclaw University of Technology, Poland |
| Frati, Fulvio | University of Milan, Italy |
| Fuchino, Tetsuo | Tokyo Institute of Technology, Japan |
| Fujimoto, Taro | Fujitsu Laboratories Limited, Japan |
| Fujinami, Tsutomu | Japan Advanced Institute of Science and Technology, Japan |
| Fukue, Yoshinori | Fujitsu Laboratories Limited, Japan |
| Gallego-Merino, Miren Josune | Universidad del Pais Vasco, Spain |
| Gamberger, Dragan | Rudjer Boskovic Institute, Croatia |
| Gao, Kun | Zhejiang Wanli University, China |
| Gao, Ying | Saitama University, Japan |
| Garcia-Sebastian, Maite | Universidad del Pais Vasco, Spain |
| Gastaldo, Paolo | University of Genoa, Italy |
| Gendarmi, Domenico | University of Bari, Italy |
| Georgieva, Petia | University of Aveiro, Portugal |
| Gianfelici, Francesco | Università Politecnica delle Marche, Italy |
| Gianini, Gabriele | University of Milan, Italy |
| Giordano, Roberto | Federal University of Sao Carlos, Brazil |
| Giorgini, Paolo | University of Trento, Italy |
| Gledec, Gordan | University of Zagreb, Croatia |
| Gold, Hrvoje | University of Zagreb, Croatia |
| Goldstein, Pavle | University of Zagreb, Croatia |
| Gomez-Dans, Jose Luis | University College London, UK |
| Goto, Yuichi | Saitama University, Japan |
| Grana, Manuel | Universidad del Pais Vasco, Spain |
| Greenwood, Garrison | Portland State University, USA |
| Gu, Dongbing | University of Essex, UK |
| Guo, Huawei | Shanghai Jiao Tong University, China |
| Hakansson, Anne | Uppsala University, Sweden |
| Halilcevic, Suad | University of Tuzla, Bosnia & Herzegovina |
| Hammer, Barbara | Clausthal University of Technology, Germany |
| Hanachi, Chihab | University Toulouse 1 and IRIT Laboratory, France |
| Hara, Akira | Hiroshima City University, Japan |
| Harada, Koji | Toyohashi University of Technology, Japan |
| Harris, Irina | Cardiff University, UK |
| Harris, Richard | University of Lancaster, UK |
| Hartung, Ronald L. | Franklin University, USA |
| Hasegawa, Shinobu | Japan Advanced Institute of Science and Technology, Japan |
| Hayashi, Hidehiko | Naruto University of Education, Japan |
| Hernandez, Carmen | Universidad del Pais Vasco, Spain |

| | |
|---|---|
| Herrero, Alvaro | University of Burgos, Spain |
| Hildebrand, Lars | Technical University of Dortmund, Germany |
| Hiroshi, Mineno | Shizuoka University, Japan |
| Handa, Hisashi | Okayama University, Japan |
| Hocenski, Zeljko | University of Osijek, Croatia |
| Hori, Satoshi | Monotsukuri Institute of Technologists, Japan |
| Howlett, Robert J. | University of Brighton, UK |
| Hruschka, Eduardo | University of Sao Paulo, Brazil |
| Huang, Xu | University of Canberra, Australia |
| Huljenic, Darko | Ericsson Nikola Tesla, Croatia |
| Ichimura, Takumi | Hiroshima City University, Japan |
| Inuzuka, Nobuhiro | Nagoya Institute of Technology, Japan |
| Ioannidis, Stratos | University of the Aegean, Greece |
| Ishibuchi, Hisao | Osaka Prefecture University, Japan |
| Ishida, Yoshiteru | Toyohashi University of Technology, Japan |
| Ishii, Naohiro | Aichi Institute of Technology, Japan |
| Ito, Hideaki | Chukyo University, Japan |
| Ito, Kazunari | Aoyama University, Japan |
| Ito, Sadanori | Tokio University of Agriculture and Technology, Japan |
| Itou, Junko | Wakayama University, Japan |
| Iwahori, Yuji | Chubu University, Japan |
| Jacquenet, Francois | University of Saint-Etienne, France |
| Jain, Lakhmi C. | University of South Australia, Australia |
| Jarman, Ian | Liverpool John Moores University, UK |
| Jatowt, Adam | Kyoto University, Japan |
| Jevtic, Dragan | University of Zagreb, Croatia |
| Jezic, Gordan | University of Zagreb, Croatia |
| Jiang, Jianmin | University of Bradford, UK |
| Jimenez-Berni, Jose Antonio | IAS-CSIC, Spain |
| Johnson, Ray | Defence Science and Technology Organisation, Australia |
| Ju, Zhaojie | University of Portsmouth, UK |
| Jung, Jason | Yeungnam University, Korea |
| Juszczyszyn, Krzysztof | Wroclaw University of Technology, Poland |
| Kaczmarek, Tomasz | Poznan University of Economics, Poland |
| Karny, Miroslav | Academy of Sciences of the Czech Republic, Czech Republic |
| Karwowski, Waldemar | University of Central Florida, USA |
| Katarzyniak, Radoslaw | Wrocaw University of Technology, Poland |
| Kato, Shohei | Nagoya Institute of Technology, Japan |
| Katsifarakis, Konstantinos | Aristotelian University of Thessaloniki, Greece |
| Kazienko, Przemyslaw | Wroclaw University of Technology, Poland |
| Kecman, Vojislav | University of Auckland, New Zealand |
| Keysers, Daniel | Google Switzerland, Switzerland |

| | |
|---|---|
| Kim, Dongwon | Korea University, Korea |
| Kimura, Masahiro | Ryukoku University, Japan |
| Kojiri, Tomoko | Nagoya University, Japan |
| Kolaczek, Grzegorz | Wroclaw University of Technology, Poland |
| Koukam, Abder | Université de Technologie de Belfort Montbeliard, France |
| Kovacic, Zdenko | University of Zagreb, Croatia |
| Krajnovic, Sinisa | Nippon Ericsson K.K., Japan |
| Krol, Dariusz | Wroclaw University of Technology, Poland |
| Kubota, Naoyuki | Tokyo Metropolitan University, Japan |
| Kucheryavskiy, Sergey | Aalborg University Esbjerg, Denmark |
| Kume, Terunobu | Fujitsu Laboratories Limited, Japan |
| Kunifuji, Susumu | Japan Advanced Institute of Science and Technology, Japan |
| Kuroda, Chiaki | Tokyo Institute of Technology, Japan |
| Kusek, Mario | University of Zagreb, Croatia |
| Lanjeri, Siham | University of Cordoba, Spain |
| Le, Kim | University of Canberra, Australia |
| Lee, Hsuan-Shih | National Taiwan Ocean University, Taiwan |
| Lee, Huey-Ming | Chinese Culture University, Taiwan |
| Lee, Malrey | ChonBuk National University, Korea |
| Lennox, Barry | University of Manchester, UK |
| Leray, Philippe | University of Nantes, France |
| Lhotska, Lenka | Czech Technical University, Czech Republic |
| Lisboa, Paulo | Liverpool John Moores University, UK |
| Liu, Honghai | University of Portsmouth, UK |
| Liu, Yonghuai | Aberystwyth University, UK |
| Loia, Vincenzo | University of Salerno, Italy |
| Lonetti, Francesca | University of Pisa, Italy |
| Loo, Chu-Kiong | Multimedia University, Malaysia |
| López, Beatriz | Universitat de Girona, Spain |
| Lops, Pasquale | University of Bari, Italy |
| Lovrek, Ignac | University of Zagreb, Croatia |
| Ma, Wanli | University of Canberra, Australia |
| MacDonald, Bruce | University of Auckland, New Zealand |
| Magnani, Lorenzo | University of Pavia, Italy |
| Malchiodi, Dario | University of Milan, Italy |
| Maldonado-Bautista, Jose O. | Universidad del Pais Vasco, Spain |
| Maojo, Victor | Universidad Politecnica de Madrid, Spain |
| Maratea, Antonio | University of Naples Parthenope, Italy |
| Margaris, Dionissios | University of Patras, Greece |
| Marinai, Simone | University of Florence, Italy |
| Markovic, Hrvoje | Tokyo Institute of Technology, Japan |
| Marrara, Stefania | University of Milan, Italy |
| Martin, Luis | Universidad Politecnica de Madrid, Spain |

| | |
|---|---|
| Martin-Sanchez, Fernando | Institute of Health Carlos III, Spain |
| Masulli, Francesco | University of Genoa, Italy |
| Matijasevic, Maja | University of Zagreb, Croatia |
| Matsuda, Noriyuki | Wakayama University, Japan |
| Matsudaira, Kazuya | Shizuoka University, Japan |
| Matsui, Nobuyuki | University of Hyogo, Japan |
| Matsumoto, Hideyuki | Tokyo Institute of Technology, Japan |
| Matsushita, Mitsunori | NTT Communication Science Labs, Japan |
| McCormac, Andrew | Alpha Data Ltd, UK |
| Mencar, Corrado | University of Bari, Italy |
| Meng, Qinggang | University of Loughborough, UK |
| Menolascina, Filippo | Technical University of Bari, Italy |
| Mera, Kazuya | Hiroshima City University, Japan |
| Minazuki, Akinori | Kushiro Public University of Economics, Japan |
| Mineno, Hiroshi | Shizuoka University, Japan |
| Minoru, Fukumi | Tokushima University, Japan |
| Misue, Kazuo | University of Tsukuba, Japan |
| Mitsukura, Yasue | Tokyo University of Agriculture and Technology, Japan |
| Mituhara, Hiroyuki | Tokushima University, Japan |
| Miura, Hirokazu | Wakayama University, Japan |
| Miura, Motoki | Japan Advanced Institute of Science and Technology, Japan |
| Miyadera, Youzou | Tokyo Gakugei University, Japan |
| Mizuno, Tadanori | Shizuoka University, Japan |
| Mohammadian, Masoud | University of Canberra, Australia |
| Moraga, Claudio | University of Dortmund, Germany |
| Mukai, Naoto | Tokyo University of Science, Japan |
| Mumford, Christine | Cardiff University, UK |
| Munemori, Jun | Wakayama University, Japan |
| Nachtegael, Mike | Ghent University, Belgium |
| Nakada, Toyohisa | Japan Advanced Institute of Science and Technology, Japan |
| Nakamatsu, Kazumi | University of Hyogo, Japan |
| Nakamura, Tsuyoshi | Nagoya Institute of Technology, Japan |
| Nakano, Ryohei | Nagoya Institute of Technology, Japan |
| Nakao, Zensho | University of the Ryukyus, Japan |
| Napolitano, Francesco | University of Salerno, Italy |
| Nara, Shinsuke | Saitama University, Japan |
| Nara, Yumiko | The Open University of Japan, Japan |
| Nascimiento, Jose | Instituto Superior de Engenharia de Lisboa, Portugal |
| Nebot, Angela | Technical University of Catalonia, Spain |
| Negoita, Mircea Gh. | KES International |

| | |
|---|---|
| Ng, Wilfred | Hong Kong University of Science and Technology, China |
| Nguyen, Ngoc Thanh | Wroclaw University of Technology, Poland |
| Nicolau, Viorel | Dunarea de Jos University of Galati, Romania |
| Nicoletti, Maria do Carmo | Federal University of Sao Carlos, Brazil |
| Nicosia, Giuseppe | University of Catania, Italy |
| Nijholt, Anton | University of Twente, The Netherlands |
| Nishida, Toyoaki | Kyoto University, Japan |
| Nishimoto, Kazushi | Japan Advanced Institute of Science and Technology |
| Nobuhara, Hajime | University of Tsukuba, Japan |
| Nowe, Ann | VUB, Belgium |
| Nowostawski, Mariusz | University of Otago, New Zealand |
| O'Grady, Michael | University College Dublin, Ireland |
| Okamoto, Takeshi | Kanagawa Institute of Technology, Japan |
| Oltean, Gabriel | Technical University of Cluj-Napoca, Romania |
| Ortega, Juan | Universidad de Sevilla, Spain |
| Ozawa, Seiichi | Kobe University, Japan |
| Palade, Vasile | University of Oxford, UK |
| Pan, Dan China | Mobile Group Guangdong Branch, China |
| Pan, Jeng-Shyang | National Kaohsiung University of Applied Sciences, Taiwan |
| Pandzic, Igor S. | University of Zagreb, Croatia |
| Papathanassiou, Stavros | National Technical University of Athens, Greece |
| Park, Gwi-Tae | Korea University, Korea |
| Parra-Llanas, Xavier | Technical University of Catalonia, Spain |
| Pasero, Eros | Politecnico di Torino, Italy |
| Paz, Abel Francisco | University of Extremadura, Spain |
| Pedrycz, Witold | University of Alberta, Canada |
| Pehcevski, Jovan | MIT Faculty of Information Technologies, Macedonia |
| Perez del Rey, David | Universidad Politecnica de Madrid, Spain |
| Perez, Rosa M. | University of Extremadura, Spain |
| Perez-Lopez, Carlos | Technical University of Catalonia, Spain |
| Pessa, Eliano | University of Pavia, Italy |
| Pham, Tuan | James Cook University, Australia |
| Phillips, Phil | Office for National Statistics, UK |
| Phillips-Wren, Gloria | Loyola College in Maryland, USA |
| Picasso, Francesco | University of Genoa, Italy |
| Pieczynska, Agnieszka | Instytut Informatyki Technicznej, Poland |
| Pirrone, Roberto | University of Palermo, Italy |
| Plaza, Antonio | University of Extremadura, Spain |
| Popa, Rustem | Dunarea de Jos University of Galati, Romania |
| Popescu, Daniela | University of Oradea, Romania |

| | |
|---|---|
| Ragone, Azzurra | Technical University of Bari, Italy |
| Raiconi, Giancarlo | University of Salerno, Italy |
| Raimondo, Giovanni | Politecnico di Torino, Italy |
| Ramon, Jan | Katholieke Universiteit Leuven, Belgium |
| Ranawana, Romesh | ClearView Scientific, UK |
| Razmerita, Liana | Copenhagen Business School, Denmark |
| Reghunadhan, Rajesh | Bharathiar University, India |
| Remagnino, Paolo | Kingston University, UK |
| Resta, Marina | University of Genoa, Italy |
| Reusch, Bernd | Technical University of Dortmund, Germany |
| Rizzo, Donna | University of Vermont, USA |
| Rohani, Bakar | Waseda University, Japan |
| Romero, Enrique | Technical University of Catalonia, Spain |
| Rosic, Marko | University of Split, Croatia |
| Rovas, Dimitrios | Technical University of Crete, Greece |
| Rozic, Nikola | University of Split, Croatia |
| Saito, Kazumi | University of Shizuoka, Japan |
| Sakai, Hiroshi | Kyushu Institute of Technology, Japan |
| Sakamoto, Ryuuki | Wakayama University, Japan |
| Sanchez, Eduardo | Logic Systems Laboratory IN-Ecublens, Switzerland |
| Sassi, Roberto | University of Milan, Italy |
| Sato-Ilic, Mika | University of Tsukuba, Japan |
| Scarselli, Franco | University of Siena, Italy |
| Schanda, Janos | University of Veszprem, Hungary |
| Schwenker, Friedhelm | University of Ulm, Germany |
| Sebillot, Pascale | IRISA/INSA de Rennes, France |
| Semeraro, Giovanni | University of Bari, Italy |
| Sergiadis, Georgios | Aristotelian University of Thessaloniki, Greece |
| Serra-Sagrista, Joan | Universitat Autonoma Barcelona, Spain |
| Sharma, Dharmendra | University of Canberra, Australia |
| Shiau, Yea-Jou | China University of Technology, Taiwan |
| Shin, Jungpil | University of Aizu, Japan |
| Shinagawa, Norihide | Tokyo University of Agriculture and Technology, Japan |
| Shizuki, Buntarou | University of Tsukuba, Japan |
| Shkodirev, Viacheslaw | St. Petersburg State Polytechnic University, Russia |
| Sidhu, Amandeep | Curtin University of Technology, Australia |
| Signore, Oreste | Istituto di Scienza e Tecnologie dell' Informazione A. Faedo, Italy |
| Sikic, Mile | University of Zagreb, Croatia |
| Sinkovic, Vjekoslav | University of Zagreb, Croatia |
| Smuc, Tomislav | Rudjer Boskovic Institute, Croatia |
| Snajder, Jan | University of Zagreb, Croatia |

| | |
|---|---|
| Sobecki, Janusz | Wroclaw University of Technology, Poland |
| Sohn, Surgwon | Hoseo University, Korea |
| Somol, Petr | Institute of Information Theory and Automation, Czech Republic |
| Staiano, Antonino | University of Naples Federico II, Italy |
| Stamou, Giorgos | National Technical University of Athens, Greece |
| Stankov, Slavomir | University of Split, Croatia |
| Stecher, Rodolfo | L3S Research Center, Germany |
| Stellato, Armando | University of Rome, Italy |
| Stoermer, Heiko | University of Trento, Italy |
| Strahil, Ristov | Rudjer Boskovic Institute, Croatia |
| Sugihara, Taro | Japan Advanced Institute of Science and Technology, Japan |
| Sugiyama, Kozo | Japan Advanced Institute of Science and Technology, Japan |
| Sulaiman, Ross | University of Canberra, Australia |
| Supek, Fran | Rudjer Boskovic Institute, Croatia |
| Surjan, Gyoergy | National Institute for Strategic Health Research, Hungary |
| Suzuki, Nobuo | KDDI Corporation, Japan |
| Tabakow, Iwan | Wroclaw University of Technology, Poland |
| Tadanori, Mizuno | Shizuoka University, Japan |
| Tadic, Marko | University of Zagreb, Croatia |
| Tagawa, Takahiro | Kyushu University, Japan |
| Tagliaferri, Roberto | University of Salerno, Italy |
| Takahashi, Osamu | Future University-Hakodate, Japan |
| Takahashi, Shin | University of Tsukuba, Japan |
| Takeda, Kazuhiro | Shizuoka University, Japan |
| Takenaka, Tomoya | Shizuoka University, Japan |
| Taki, Hirokazu | Wakayama University, Japan |
| Tamani, Karim | University of Savoie, France |
| Tanahashi, Yusuke | Nagoya Institute of Technology, Japan |
| Tateiwa, Yuichiro | Nagoya University, Japan |
| Tesar, Ludvik | Academy of Sciences of the Czech Republic, Czech Republic |
| Thai, Hien | University of the Ryukyus, Japan |
| Ting, Hua Nong | University of Malaya, Malaysia |
| Ting, I-Hsien | National University of Kaohsiung, Taiwan |
| Tohru, Matsuodani | Debag Engineering Ltd, Japan |
| Tonazzini, Anna | Istituto di Scienza e Tecnologie dell' Informazione A. Faedo, Italy |
| Torsello, Maria Alessandra | University of Bari, Italy |
| Tran, Dat | University of Canberra, Australia |
| Trentin, Edmondo | University of Siena, Italy |

| | |
|---|---|
| Trzec, Krunoslav | Ericsson Nikola Tesla, Croatia |
| Tsourveloudis, Nikos | Technical University of Crete, Greece |
| Tsumoto, Shusaku | Shimane University, Japan |
| Turchetti, Claudio | Università Politecnica delle Marche, Italy |
| Tweedale, Jeffrey | Defence Science and Technology Organisation, Australia |
| Uchino, Eiji | Yamaguchi University, Japan |
| Ugai, Takanori | Fujitsu Laboratories Limited, Japan |
| Ushiama, Teketoshi | Kyushu University, Japan |
| Vaklieva-Bancheva, Natasha | Bulgarian Academy of Sciences, Bulgaria |
| Vassanyi, Istvan | University of Pannonia, Hungary |
| Vega, Miguel | University of Granada, Spain |
| Veganzones, Miguel Angel | Universidad del Pais Vasco, Spain |
| Vellido, Alfredo | Technical University of Catalonia, Spain |
| Vitabile, Salvatore | University of Palermo, Italy |
| Vohland, Michael | Trier University, Germany |
| Wang, Jin-Long | Ming Chuan University, Taiwan |
| Wang, Yang | University of Portsmouth, UK |
| Watada, Junzo | Waseda University, Japan |
| Watanabe, Toyohide | Nagoya University, Japan |
| Watanabe, Yuji | Nagoya City University, Japan |
| Weber, Cornelius | Frankfurt Institute for Advanced Studies, Germany |
| Whitaker, Roger | Cardiff University, UK |
| Xia, Feng | Queensland University of Technology, Australia |
| Yamada, Kunihiro | Tokai University, Japan |
| Yamashita, Yoshiyuki | Tokyo University of Agriculture and Technology, Japan |
| Yasuda, Takami | Nagoya University, Japan |
| Yip, Chi Lap | University of Hong Kong, China |
| Yi-Sheng, Huang | Chung Cheng Institute of Technology, Taiwan |
| Yoshida, Kenichi | University of Tsukuba, Japan |
| Yoshida, Kouji | Shonan Institute of Technology, Japan |
| Yoshiura, Noriaki | Saitama University, Japan |
| Younan, Nick | Mississippi State University, USA |
| Yu, Donggang | James Cook University, Australia |
| Yu, Zhiwen | Kyoto University, Japan |
| Yuizono, Takaya | Japan Advanced Institute of Science and Technology, Japan |
| Yukawa, Takashi | Nagaoka University of Technology, Japan |
| Zalili, Musa | Waseda University, Japan |
| Zare, Alina | University of Florida, USA |
| Zebulum, Ricardo | NASA Jet Propulsion Laboratory, USA |
| Zeng, An | Guangdong University of Technology, China |
| Zeng, Xiangyan | University of California Davis, USA |

| | |
|---|---|
| Zhang, Bailing | Xi'an Jiaotong-Liverpool University, China |
| Zhu, Goupu | Sun Yat-Sen University, China |
| Zippo, Antonio | University of Milan, Italy |
| Zunino, Rodolfo | University of Genoa, Italy |

## General Track Chairs

Artificial Neural Networks and Connectionist Systems
Bruno Apolloni, University of Milan, Italy

Fuzzy and Neuro–Fuzzy Systems
Bernd Reusch, Technical University of Dortmund, Germany

Evolutionary Computation
Zensho Nakao, University of the Ryukyus, Japan

Machine Learning and Classical AI
Floriana Esposito, University of Bari, Italy

Agent Systems
Ngoc Thanh Nguyen, Wroclaw University of Technology, Poland

Knowledge Based and Expert Systems
Anne Håkansson, Uppsala University, Sweden

Hybrid Intelligent Systems
Vasile Palade, University of Oxford, UK

Intelligent Vision and Image Processing
Tuan Pham, James Cook University, Australia

Knowledge Management, Ontologies and Data Mining
Bojana Dalbelo Basic, University of Zagreb, Croatia

Web Intelligence, Text and Multimedia Mining and Retrieval
Andreas Nuernberger, University of Magdeburg, Germany

Intelligent Signal Processing
Miroslav Karny, Academy of Sciences of the Czech Republic, Czech Republic

Intelligent Robotics and Control
Honghai Liu, University of Portsmouth, UK

## Invited Session Chairs

Advanced Groupware
Jun Munemori (Wakayama University, Japan), Hiroshi Mineno (Shizuoka University, Japan)

Advanced Knowledge-Based Systems
Alfredo Cuzzocrea (University of Calabria, Italy)

Advanced Neural Processing Systems
Monica Bianchini, Marco Maggini, Franco Scarselli (Università degli Studi di Siena, Italy)

Agent and Multi-Agent Systems: Technologies and Applications
Bala M. Balachandran, Dharmendra Sharma (University of Canberra, Australia)

Ambient Intelligence
Cecilio Angulo (Technical University of Catalonia, Spain), Honghai Liu (University of Portsmouth, UK)

Application of Knowledge Models in Healthcare
István Vassányi (University of Pannonia, Hungary), Gyoergy Surjan (National Institute for Strategic Health Research, Hungary)

Artificial Intelligence Driven Engineering Design Optimization
Ioannis K. Nikolos (Technical University of Crete, Greece)

Biomedical Informatics: Intelligent Information Management from Nanomedicine to Public Health
Victor Maojo (Universidad Politecnica de Madrid, Spain)

Chance Discovery
Akinori Abe (ATR Knowledge Science Laboratories, Japan), Yukio Ohsawa (University of Tokyo, Japan)

Communicative Intelligence
Ngoc Thanh Nguyen (Wroclaw University of Technology, Poland), Toyoaki Nishida (Kyoto University, Japan)

Computational Intelligence for Image Processing and Pattern Recognition
Yen-Wei Chen (Ritsumeikan University, Japan)

Computational Intelligence in Human Cancer Research
Alfredo Vellido (Technical University of Catalonia, Spain), Paulo J.G. Lisboa (Liverpool John Moores University, UK)

Computational Intelligence Techniques for Knowledge Discovery
Claudio Turchetti, Paolo Crippa, Francesco Gianfelici (Università Politecnica delle Marche, Italy)

Computational Intelligence Techniques for Web Personalization
Giovanna Castellano, Alessandra Torsello (University of Bari, Italy)

Computational Intelligent Techniques for Bioprocess Modelling, Monitoring and Control
Maria do Carmo Nicoletti, Teresa Cristina Zangirolami, Estevam Rafael Hruschka Jr. (Federal University of Sao Carlos, Brazil)

Contributions of Intelligent Decision Technologies (IDT)
Gloria Phillips-Wren (Loyola College in Maryland, USA), Lakhmi C. Jain (University of South Australia, Australia)

Engineered Applications of Semantic Web – SWEA
Tommaso Di Noia, Marco Degemmis, Giovanni Semeraro, Eugenio Di Sciascio (University of Bari, Italy)

Enhance Secure User Authentication Through Intelligent and Strong Techniques
Ernesto Damiani, Antonia Azzini (University of Milan, Italy)

Evolutionary Multiobjective Optimization
Christine Mumford (Cardiff University, UK)

Evolvable Hardware and Adaptive Systems – Advanced Engineering Design Methodologies and Applications
Mircea Gh. Negoita (KES International), Sorin Hintea (Technical University of Cluj-Napoca, Romania)

Evolvable Hardware Applications in the Area of Electronic Circuits Design
Mircea Gh. Negoita (KES International), Sorin Hintea (Technical University of Cluj-Napoca, Romania)

Hyperspectral Imagery for Remote Sensing: Intelligent Analysis and Applications
Miguel A. Veganzones, Manuel Grana (Universidad del Pais Vasco, Spain)

Immunity-Based Systems
Yoshiteru Ishida (Toyohashi University of Technology, Japan)

Innovation-oriented Knowledge Management Platform
Toyohide Watanabe (Nagoya University, Japan), Taketoshi Ushiama (Kyushu University, Japan)

Innovations in Intelligent Multimedia Systems
Cecilia Sik Lanyi (University of Pannonia, Hungary), Lakhmi C. Jain (University of South Australia, Australia)

Innovations in Virtual Reality
Cecilia Sik Lanyi (University of Pannonia, Hungary), Lakhmi C. Jain (University of South Australia, Australia)

Intelligent Computing for Grid
Kun Gao (Zhejiang Wanli University, China)

Intelligent Data Processing in Process Systems and Plants
Yoshiyuki Yamashita (Tokyo University of Agriculture and Technology, Japan), Tetsuo Fuchino (Tokyo Institute of Technology, Japan)

Intelligent Environment Support for Collaborative Learning
Toyohide Watanabe, Tomoko Kojiri (Nagoya University, Japan)

Intelligent Systems in Medicine and Healthcare
Chee-Peng Lim, (University of Science Malaysia, Malaysia), Lakhmi C. Jain (University of South Australia, Australia), Robert F. Harrison (University of Sheffield, UK)

Intelligent Systems in Medicine: Innovations in Computer–Aided Diagnosis and Treatment
Mia Markey (University of Texas at Austin, USA), Lakhmi C. Jain (University of South Australia, Australia)

Intelligent Utilization of Soft Computing Techniques
Norio Baba (Osaka Kyoiku University, Japan)

Knowledge-Based Interface Systems [I]
Naohiro Ishii (Aichi Institute of Technology, Japan), Yuji Iwahori (Chubu University, Japan)

Knowledge-Based Interface Systems [II]
Yoshinori Adachi (Chubu University, Japan), Nobuhiro Inuzuka (Nagoya Institute of Technology, Japan)

Knowledge Interaction for Creative Learning
Toyohide Watanabe, Tomoko Kojiri (Nagoya University, Japan)

Knowledge-Based Creativity Support Systems
Susumu Kunifuji (Japan Advanced Institute of Science and Technology, Japan), Kazuo Misue (University of Tsukuba, Japan), Motoki Miura (Japan Advanced Institute of Science and Technology, Japan), Takanori Ugai (Fujitsu Laboratories Limited, Japan)

Knowledge-Based Multi-criteria Decision Support
Hsuan-Shih Lee (National Taiwan Ocean University, Taiwan)

Knowledge-Based Systems for e-Business
Kazuhiko Tsuda (University of Tsukuba, Japan)

Neural Information Processing for Data Mining
Ryohei Nakano (Nagoya Institute of Technology, Japan), Kazumi Saito (University of Shizuoka, Japan)

Neural Networks in Image Processing
Monica Bianchini, Marco Maggini, Franco Scarselli (Università degli Studi di Siena, Italy)

New Advances in Defence and Security Systems in Intelligent Environments
Jadranka Sunde (Defence Science and Technology Organisation, Australia), Lakhmi C. Jain (University of South Australia, Australia)

Novel Foundation and Applications of Intelligent Systems
Valentina E. Balas (Aurel Vlaicu University of Arad, Romania), Chee-Peng Lim, (University of Science Malaysia, Malaysia), Lakhmi C. Jain (University of South Australia, Australia)

Reasoning-Based Intelligent Systems
Kazumi Nakamatsu (University of Hyogo, Japan)

Relevant Reasoning for Discovery and Prediction
Jingde Cheng, Yuichi Goto (Saitama University, Japan)

Skill Acquisition and Ubiquitous Human Computer Interaction
Hirokazu Taki (Wakayama University, Japan), Satoshi Hori (Monotsukuri Institute of Technologists, Japan)

Soft Computing Approach to Management Engineering
Junzo Watada (Waseda University, Japan), Huey-Ming Lee (Chinese Cultural University, Taiwan), Taki Kanda (Bunri University of Hospitality, Japan)

Smart Sustainability
Robert J. Howlett (University of Brighton, UK)

Spatio-temporal Database Concept Support for Organizing Virtual Earth
Toyohide Watanabe (Nagoya University, Japan), Jun Feng (Hohai University, Japan), Naoto Mukai (Tokyo University of Science, Japan)

Unsupervised Clustering for Exploratory Data Analysis
Roberto Tagliaferri (University of Salerno, Italy), Michele Ceccarelli (University of Sannio, Italy)

Use of AI Techniques to Build Enterprise Systems
Ronald L. Hartung (Franklin University, USA)

XML Security
Ernesto Damiani, Stefania Marrara (University of Milan, Italy)

3D Approaches for Visual Facial Expression and Emotion Dynamics Recognition in a Real Time Context
Andoni Beristain, Manuel Grana (Universidad del Pais Vasco, Spain)

## Sponsoring Institutions

Ministry of Science, Education and Sports of the Republic of Croatia
University of Zagreb, Faculty of Electrical Engineering and Computing
Ericsson Nikola Tesla, Zagreb, Croatia
Croatian National Tourist Board
Zagreb Tourist Board

# Table of Contents – Part II

## II Intelligence Everywhere

## Artificial Intelligence Driven Engineering Design Optimization

## Biomedical Informatics: Intelligent Information Management from Nanomedicine to Public Health

## Communicative Intelligence

## Computational Intelligence for Image Processing and Pattern Recognition

## Computational Intelligence in Human Cancer Research

## Computational Intelligence Techniques for Web Personalization

## Computational Intelligent Techniques for Bioprocess Modelling, Monitoring and Control

## Intelligent Computing for Grid

## Intelligent Security Techniques

## Intelligent Utilization of Soft Computing Techniques

## Reasoning-Based Intelligent Systems

## Relevant Reasoning for Discovery and Prediction

## Spatio-Temporal Database Concept Support for Organizing Virtual Earth

# III Knowledge Everywhere

## Advanced Knowledge-Based Systems

## Chance Discovery

## Innovation-Oriented Knowledge Management Platform

## Knowledge-Based Creativity Support Systems

## Knowledge-Based Interface Systems [I]

## Knowledge-Based Interface Systems [II]

## Knowledge-Based Multi-criteria Decision Support

## Knowledge-Based Systems for e-Business

# Design of Row-Based Flexible Manufacturing System with Evolutionary Computation

Mirko Ficko and Joze Balic

University of Maribor, Faculty of mechanical engineering, Smetanova ulica 17, SI-2000
Maribor, Slovenia
{mirko.ficko,joze.balic}@uni-mb.si

**Abstract.** This paper discusses design of flexible manufacturing systems (FMSs) in one or multiple rows. Evolutionary computation, particularly genetic algorithms (GAs) proved to be successful in search of optimal solution for this type of problems. The model of solution, the most suitable way of coding the solutions into the organisms and the selected evolutionary and genetic operators are presented. In this connection, the most favourable number of rows and the sequence of devices in the individual row are established by means of genetic algorithms (GAs). In the end the test results of the application made and the analysis are discussed.

**Keywords:** flexible manufacturing system, genetic algorithms, design.

## 1 Introduction

Some years ago, when the concept of the FMS appeared for the first time, it was commercially successful in spite of successful applications. The idea was revolutionary and slightly ahead of time. At that time the production companies were not yet in position to apply organizationally and technically the FMS to their class of products. While the FMS almost ceased to be spoken of, the development went on and the idea reappeared this time as a consequence of evolution of hardware and software. Moreover, when designing the FMS the material, tool and information flows are met, which must be controlled best possible so that the system can ensure highly efficient operation. This paper is concerned about searching for optimum placing of devices and machines that the handling costs of the material, semi-finished products and products within the system are lowest possible.

The first section of the paper presents the problem of design the FMS. The second section introduces the FMS and its specific properties with respect to transport and design. The third section briefs the reader on the model proposed and on the GA method used in our work. The fourth section summarizes the results obtained by the model. The discussion and the concluding findings follow.

## 2 Problem of Layout of Devices in FMS

Optimum arrangement of devices and machines in the FMS is one of the basic requirements in design process of the FMS. Good design of such system is a basis for

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 1–8, 2008.

its efficient operation and for low operating costs [1]. This is especially true in the case of large FMS [2]. The manner of arranging of working devices largely depends on the type of production [3] and the used transport system. It was estimated that 20% to 50% of the manufacturing costs are due to handling of work pieces; by a good arrangement of devices it is possible to reduce the manufacturing costs for 10% to 30% [4]. Some other authors report even higher percentage of material handling based costs, for example Chiang and Kouvelis report that 30-70% of total manufacturing costs may be attributed to materials handling and layout [5]. Therefore, already in an early stage of designing of the FMS it is necessary to have an idea of the efficient layout of the devices.

However, FMS have somewhat different requirements for transport than the other manufacturing systems. Heragu and Kusiak pointed out the fact that the FMS differ from the conventional production systems [6]; the FMS include the devices and the machines which usually do not have identical dimensions and also the distances between the individual devices are not firmly determined [6]. Therefore it is not possible to determine in advance the locations and then to place the devices on them. Due to those facts it is practically impossible in the FMS to locate the working device and workplaces by methods based on the principle "one place one device". Design of FMS is problem of arrangement of unequally large devices. Therefore, only the methods for arranging differently large devices can be used. Generally, unequal-area layout problems are more difficult to solve than equal-area layout problems, primarily because unequal-area layout problems introduce additional constraints into the problem formulation [7].

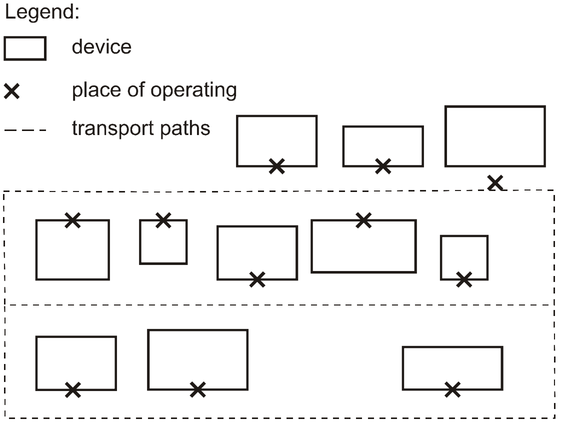


**Fig. 1.** Layout of devices fed by AGVs in multiple rows

In the initial stage of the development of the FMS the sequential transport devices were used as a result of the strong influence of conventional transfer lines [1]. However later on it was established that for greater flexibility it is favourable to use freely guided automatic vehicles. Machines fed by such transport system and machines fed by crane robot are most commonly placed in one or multiple rows (Fig. 1). The presented design method will be carried out for such type of transport system.

The problem of device placement is NP-hard. NP-hard problems are unsolvable in polynomial time [3]. Applicable mathematical solutions for such type of problem do not exist. The complexity of such problems increases exponentially with the number of devices. For instance, a FMS consisting of N machines will comprise a solution space with the size N. Similar situation occurs also in scheduling of complex production system [8]. The problem is theoretically solvable also by testing all possibilities (i.e., random searching) but practical experience shows that in such manner of solving the capabilities of either the human or the computer are fast exceeded. For arranging the devices in the FMS the number of possible solutions is equal to the number of permutations of N elements. With today's computation power of modern computers it is possible to search for the optimum solution by examining the total space of solutions only for simple FMS. In case of problems of larger dimensions it is necessary to use sophisticated solving methods which, during examining the solution space, somehow limit themselves and utilize possible solutions already examined. Searching can be executed by blind strategies or by heuristic strategies. The blind searching strategies do not use information about the problem, whereas the heuristic searching strategies use additional information for determining the best searching directions. In case of searching strategies it is necessary to distinguish:

- utilization of the best solution and (known also as gradient based search)
- searching in the space of solutions (known also as direct search [9]).

The latter is the most universal method of solving problems. This method is unusable on problems where the solution space is unknown; this is not the case in designing of FMS. The only problem is huge solution space. Evolutionary computation can handle this problem. In connection with FMSs the problems of process planning were already solved by GAs [10]. The GAs contain the elements of the methods of blind searching for the solution and directed and stochastic searching and thus give compromise between the utilization and searching for solution. GAs employ random, yet directed search for locating the globally optimal solution [11].

GAs employs the vocabulary taken from the world of genetics itself, and as a result solutions refer to organisms (genotypes) of a population. Each organism represents the code of a potential solution to a problem. A further important characteristic of GAs is that they are operating on a population of potential solutions, whereas the other search methods process a single point of the search space. The typical steps required to implement GAs are encoding of feasible solutions into organisms using a representation method, evaluation of fitness function, selection strategy, setting of GAs parameters, and criteria to terminate the process.

## 3 Model Description

Determination of the layout of FMS and its evaluation is made by GA. The sequence of the machines is created by genetic operations in the first step. In the second step the actual layout with all dimensions is created with respect to the sequence determined by GA and other technological limitations. The complete procedure of forming of the FMS with GAs is divided into these main steps:

- acquisition of technological and physical data needed for designing of the FMS,
- determination of layout by GA (determination of sequence of devices and rows),

- calculation of coordinates of devices and mutual operation points,
- determination of distances between every possible pair of devices,
- calculation of value of cost function.

For such manner of solving of the problem it is necessary to know the dimensions of devices and the minimum allowable distances between all the pairs of devices. Further it is necessary to know the transport quantities between the individual devices during a certain time period. Also the variable transport costs depending on the transport means used must be known. It is also necessary to know the width of transport paths and the greatest length of the row. This information is gained from the database of technological data, which has to be made previously. It contains the data which depend on the individual components of FMS. This is the information about the devices being arranged in the FMS, information about the transport means and information about the space where the FMS will be situated. Those data are, for instance, the prices of motions of the transport devices per unit of length, the minimum required distances between the individual machines, and the overall dimension of the devices. Frequencies of motions of the transport devices depend on the products to be made by the FMS studied. Thus, the GA takes the data needed from that database. It automatically determines the layout itself by means of genetic operations and evolution.

In the first step of the GA the initial population is created at random (Fig. 2). Only correct organisms representing feasible solution are created. This initial generation enters the evolutionary loop of the GA. After evaluation of the population the selection of organisms with the roulette wheel method is made. In the next step the operations of reproduction and crossover with probability $p_r$ and $p_c$, respectively follow. The operation of mutation is executed with probability $p_m$. Thus a new population is obtained. When the GA can't improve the solution anymore the evolution is stopped and a best layout, according to fitness function is presented as solution of the problem. The layout is then evaluated by the expert with respect to the criteria not included in the cost function and in the technological database.

### 3.1 Fitness Function, Coding of Organisms, and Genetic Operations

Fitness function is in our case the sum of variable costs in a time period. For determination of value of fitness function it is necessary to know the table of the transport quantities between the individual devices $N$ in a time period [4]. Also the variable transport costs, depending on the transport means used, must be known. For example: connection between two devices in the same FMS can be performed by another transport device than between other two devices. Thus also different transport cost per unit length result. In order to find by means of these data the optimum layout of the devices $N$, it is necessary to find the minimum of the following fitness function:

$$f = \sum_{i=1}^{N} \sum_{j=1}^{N} f_{ij} \cdot c_{ij} \cdot L_{ij} \tag{1}$$

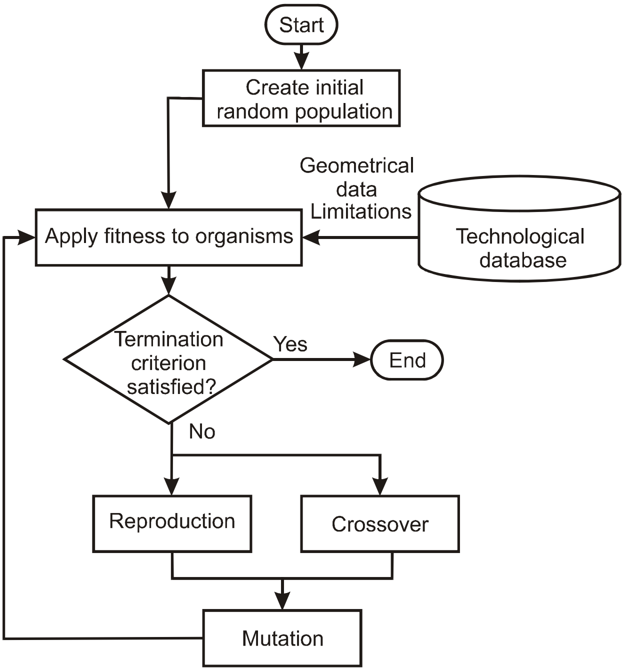


**Fig. 2.** The main steps of GA procedure

$f_{ij}$ is the frequency of trips between the devices $i$ and $j$, $c_{ij}$ are the variable transport costs for the quantity unit, and $L_{ij}$ is the length of path between the devices $i$ and $j$. The number of all devices is equal to $N$. Fitness function heavily depends on the distances $L_{ij}$ between the devices. The distance between serving points is multiplied by coefficients $f_{ij}$ and $c_{ij}$ which measure the flow and the handling cost between devices. Fitness is based on the principle that the cost of moving goes up with the distance. The fitness function itself is quite simple but it is necessary to take into account all technological and geometrical values and limitations to obtain correct value and valid solution.

Each organism represents one of the possible solutions of the problem of arranging and each gene represents one device. The most natural coding for such solution is permutation coding [12]. The sequence of genes in organism is equal to the sequence of working devices of the FMS, where the gene represents the device $i$ and its position in the organism represents the position. However, such gene would represent the arrangement in one row only. Therefore on the basis of the parameter of length of row and technological limitations the arrangement into rows is determined (Fig. 3). The number of devices in one row is limited with the maximum length of row $a$. When the length of the row is greater than $a$, the next device is placed into a new row. The procedure repeats, until all devices have been arranged into rows. Such manner of coding guarantees that all organisms are correct even after completion of genetic operations [12].

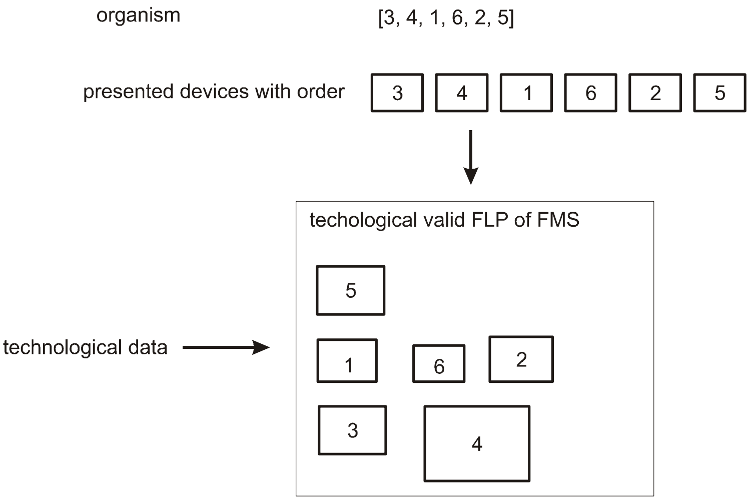


**Fig. 3.** Decoding of organisms into valid solution

The method of selection for reproduction and crossover was roulette wheel selection. In this type of selection the organisms which represent better solution have better possibility to take part in the next generation.

Many genetic operators for crossover exist for the type of coding implemented in this research. In case of crossover, two organisms are selected which are then crossed over to obtain one offspring. For the crossover the partial mapped crossover (PMX) was selected. PMX was proposed by Goldberg and Lingle [13]. PMX can be viewed as an extension of two point crossover. In addition it uses a special repairing procedure to resolve the illegitimacy caused by simple two point crossover. This type of crossover has been widely used in the field of combinatory problems. It was concluded that a PMX with a random crossing is performing better than other crossover operators to obtain a solution [14]. Proto population reached by the operation of reproduction and crossover is further modified with mutation operation. Reciprocal mutation was selected as the mutation operator. Two randomly selected genes in the original organism exchange their places. Therefore, the offspring organism represents the feasible solution. No procedure for correction of the organism is needed.

The fitness of each layout is evaluated using the above mentioned fitness function. The best organism represents the solution with lowest value of fitness function. Until the evaluation no information about the solution other than the sequence of devices is available. For the evaluation of the individual organisms, the arrangement into rows is determined. Conversion of the organism representing the sequence itself takes place as shown in Fig. 3. On this basis of the layout, the coordinates of the points of operating are determined. When calculating coordinates also the conditions and limitations from the technological database are taken into account.

## 4 Results and Conclusion

For testing of our model the test example forming of FMS with 14 devices was used. Problem of such size (*N*=14) is impossible to solve with trying all possible solutions. Therefore, our model was used with appreciation. In the first step the technological database was filled with all necessary data. In our case these were geometrical data with dimensions of devices and transport paths.

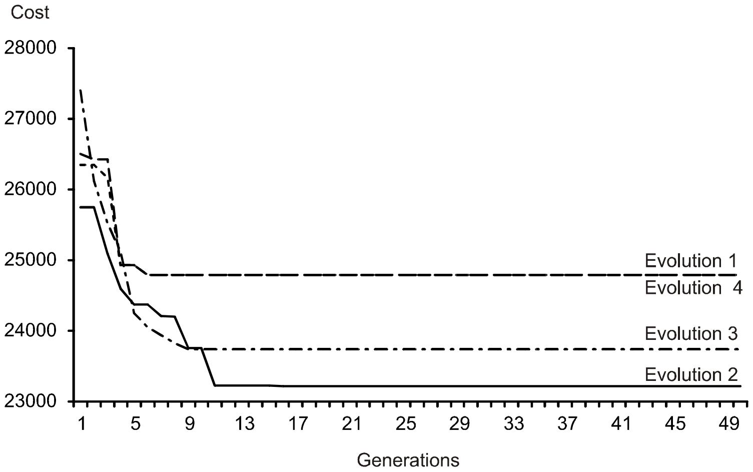


**Fig. 4.** Value of cost function during several runs of evolution

The evolution was run several times. After 4 evolutions which are presented on the Fig. 4 we have 4 different near optimal solutions. If we have a look at the solutions reached it can be found out that the model prepared similar good layouts. So the user has a possibility to choose an adequate solution from the set of high-quality solutions. The GA itself does not assure optimum solutions, but may yield near optimum solutions [15]. The proposed system can be used as a decision support tool which is used by the human expert. Usually this kind of problems is multi-criteria optimization problem, some criteria are practically impossible to add into artificial system. From this point of view it is a clear advantage to have a number of good solutions from which the expert can pick the most appropriate.

## References

1. Balic, J.: Flexible manufacturing systems, DAAAM International, Vienna (2001)
2. Hamiani, A., Popescu, G.: A knowledge-based expert system for site layout. In: Computing in Civil Engineering: Microcomputers to Supercomputers, pp. 248–256. ASCE, New York (1988)
3. Kusiak, A.: Intelligent Manufacturing Systems. Prentice-Hall, New Jersey (1990)
4. Tompkins, J.A., White, J.A., Bozer, Y.A., Frazelle, E.H., Tanchoco, J.M., Trevino, J.: Facilities Planning. John Wiley & Sons, New York (1996)

5. Chiang, W.C., Kouvelis, P.: Improved Tabu search heuristics for heuristics for solving facility layout problems. International Journal of Production Research 34(9), 2565–2585 (1996)
6. Heragu, S., Kusiak, A.: Machine layout problem in flexible manufacturing systems. Operations Research 36, 258–268 (1988)
7. Li, H., Love, E.D.: Genetic search for solving construction site-level unequal-area facility layout problems. Automation in Construction 9, 217–226 (2000)
8. Tsourveloudis, N., Ioannidis, S., Valavanis, K.: Fuzzy surplus based distributed control of manufacturing system. Advances in Production Engineering & Management 1(1), 5–12 (2006)
9. Curkovic, P., Jerebic, B.: Honey-bees optimization algorithm applied to path planning problem. International journal of simulation modelling 6(3), 154–164 (2007)
10. Viraj, T., Ajai, J.: Assessing the effectiveness of flexible process plans for loading and part type selection in FMS. Advances in Production Engineering & Management 3(1), 27–44 (2008)
11. Heng, L., Love, E.D.: Genetic search for solving construction site-level unequal-area facility layout problems. Automation in Construction 9, 217–226 (2000)
12. Gen, M., Cheng, R.: Genetic Algorithms And Engineering Design. J. Wiley & Sons, New York (1997)
13. Goldberg, D.E., Lingle, R.: Alleles, loci and the traveling salesman problem. In: Proc. 1st Conference on Genetic Algorithms, pp. 154–159 (1985)
14. Tavakkoli-Moghaddain, R., Shayan, E.: Facilities layout design by genetic algorithms. Computers industrial Engineering 35(3-4), 527–530 (1998)
15. Osman, H.M., Georgy, M.E., Ibrahim, M.E.: A hybrid CAD-based construction site layout planning system using genetic algorithms. Automation in construction 12, 749–764 (2003)

# Ant Colony System-Based Algorithm for Optimal Multi-stage Planning of Distribution Transformer Sizing

Eleftherios I. Amoiralis[1], Pavlos S. Georgilakis[1], Marina A. Tsili[2], and Antonios G. Kladas[2]

[1] Department of Production Engineering and Management, Technical University of Crete,
University Campus, Chania, Greece
eamir@tee.gr, pgeorg@dpem.tuc.gr

[2] Faculty of Electrical & Computer Engineering, National Technical University of Athens
University Campus, Athens, Greece
{mtsili, kladasel}@central.ntua.gr

**Abstract.** This paper proposes a stochastic optimization method, based on ant colony optimization, for the optimal choice of transformer sizes to be installed in a distribution network. This method is properly introduced to the solution of the optimal transformer sizing problem, taking into account the constraints imposed by the load the transformer serves throughout its life time and the possible transformer thermal overloading. The possibility to upgrade the transformer size one or more times throughout the study period results to different sizing paths, and ant colony optimization is applied in order to determine the least cost path, taking into account the transformer capital cost as well as the energy loss cost during the study period. The results of the proposed method demonstrate the benefits of its application in the distribution network planning.

**Keywords:** Transformers; Optimal Transformer Sizing; Ant Colony Optimization; Thermal Loading; Energy Loss Cost; Distribution Network Planning.

## 1 Introduction

The objective of the optimal transformer sizing problem in a multi-year planning period is to select the transformer sizes (i.e., rated capacities) and the years of transformer installation so as to serve a distribution substation load at the minimum total cost (i.e., sum of transformer purchasing cost plus transformer energy loss cost). Deterministic optimization methods may be used for the solution of this problem, such as dynamic programming [1] or integer programming [2]. However, the wide spectrum of transformer sizes and various load types involved in the electric utility distribution system make the transformer sizing a difficult combinatorial optimization problem, since the space of solutions is huge. That is why stochastic optimization methods may prove to provide more robust solutions.

In this paper, the Optimal Transformer Sizing (OTS) problem is solved by means of the heuristic Ant System method using the Elitist strategy, called Elitist Ant System (EAS). EAS belongs to the family of Ant Colony Optimization (ACO) algorithms.

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 9–17, 2008.

Dorigo has proposed the EAS in [3]. The EAS is a biologically inspired meta-heuristics method in which a colony of artificial ants cooperates in finding good solutions to difficult discrete optimization problems, such as the OTS problem. Cooperation is the key design component of ACO algorithms, i.e. allocation of the computational resources to a set of relatively simple agents (artificial ants) that communicate indirectly by stigmergy (by indirect communication mediated by the environment). In other words, a set of artificial ants cooperate in dealing with a problem by exchanging information via pheromones deposited on a graph. In the literature, ACO algorithms have been applied to solve a variety of well-known combinatorial optimization problems, such as routing [4], scheduling [5], and subset [6] problems. More details on ACO application in the solution of other problems are described in [7].

EAS was introduced in the solution of the OTS problem in case of one (three-phase, oil-immersed) distribution transformer with constant economic factors in [8], whereas this paper extends the use of EAS, taking into account all details of the economic analysis, such as the inflation rate that influences the energy loss cost and the transformer investment as well as the installation and depreciation cost in a real distribution network, constituting an efficient methodology for transformer planning. The OTS problem is solved as a constrained optimization problem.

## 2 Overview of the Proposed Method

The OTS problem consists in finding the proper capacities of transformers to be installed in a distribution network so that the overall installation and energy loss cost over the study period is minimized and the peak loading condition is met [1]. The proposed solution to the OTS problem is described in Fig. 1.

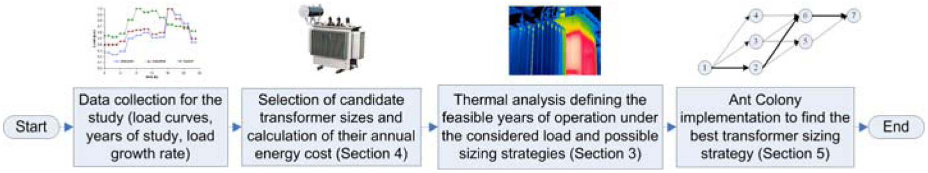

Start
Data collection for the study (load curves, years of study, load growth rate)
Selection of candidate transformer sizes and calculation of their annual energy cost (Section 4)
Thermal analysis defining the feasible years of operation under the considered load and possible sizing strategies (Section 3)
Ant Colony implementation to find the best transformer sizing strategy (Section 5)
End
1
2
3
4
5
6
7

**Fig. 1.** Flowchart illustrating the main steps of the proposed method

## 3 Calculation of Transformer Thermal Loading

The transformer thermal calculation is implemented according to the guidelines imposed by the IEEE Standard C57.91-1995 (R2002), [9]. The transformer top-oil rise over ambient temperature and winding hottest spot temperature are calculated during each interval of the considered load cycle, taking properly into account its constructional characteristics as well as the no-load and load loss. For the study of the present paper, a maximum hot-spot temperature of 120°C has been chosen, based on the relative aging rate of the insulation in the transformer. For the determination of loading limits, the calculation of the hottest spot temperature is repeated for each year of the study period, at an hourly basis, according to the daily load curve. To remain on the safe side, the peak load curve of the considered year is used in the calculations. The yearly load growth rate $s$ is taken into consideration for the derivation of the per unit load $K_t^k$ of hour $t$ at year $k$ of the study based on the per unit load $K_t^0$ of hour $t$ at year 0:

$$K_t^k = K_t^0 \cdot (1+s)^k . \qquad (1)$$

Fig. 2 shows the winding hottest spot temperature variation for six distribution transformers of rated capacity 160, 250, 300, 400, 500 and 630 kVA, serving a residential load of 398 kVA peak value at the 14th year of the study period (this load has 230 kVA peak value at the beginning of the study and 4% annual load growth). As can be observed from Fig. 2, the 160, 250 and 300 kVA transformers overcome the hottest spot limit of 120°C so they are not suitable to serve the load at the 14th year of the study period.

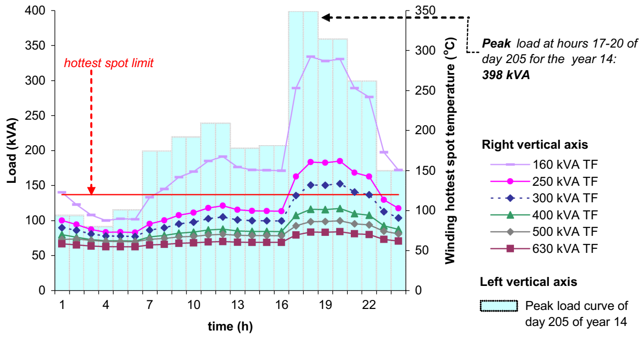


**Fig. 2.** Winding hottest spot temperature variation of six transformer (TF) ratings

## 4 Calculation of Transformer Energy Loss Cost

The calculation of annual transformer energy loss cost of the potential sizing paths is realized with the use of the energy corresponding to the transformer no-load loss (*NLL*) $E_{NLL}$ (in kWh) for each year of operation and the energy corresponding to the load loss (*LL*), $E_{LL}^k$ (in kWh) for each $k$-th year of operation. These energies are calculated according to (2) and (3), respectively:

$$E_{NLL} = NLL \cdot HPY \qquad (2), \qquad E_{LL}^k = LL \cdot \left[ l_f \cdot \frac{S_{\max,0}^l}{S_{nom}^l} (1+s)^k \right]^2 \cdot HPY \qquad (3)$$

where $S_{\max,0}^l$ is the initial peak load of the substation load type $l$ (in kVA), $S_{nom}^l$ is the nominal power of the transformer that serves load type $l$ (in kVA), *HPY* is the number of hours per year, equal to 8760, and $l_f$ is the load factor, i.e. the mean transformer loading over its lifetime (derived from the load curve of each consumer type served by the considered substation). The cost of total energy corresponding to the transformer *NLL* for each $k$-th year $C_{NLL}^k$ (in €) and the cost of energy corresponding to the transformer *LL* for each $k$-th year $C_{LL}^k$ (in €) are calculated as follows:

$$C_{NLL}^{k} = E_{NLL} \cdot CYEC^{k} \quad (4), \qquad C_{LL}^{k} = E_{LL}^{k} \cdot CYEC^{k} \quad (5)$$

where $CYEC^k$ denotes the present value of the energy cost (in €/kWh) at the $k$-th year. Finally, the total cost of the transformer energy loss $C_L^k$ for the $k$-th year is given by:

$$C_L^k = C_{NLL}^k + C_{LL}^k \,. \quad (6)$$

# 5 Elitist Ant System Method

## 5.1 Mechanism of EAS Algorithm

The EAS is an evolutionary computation optimization method based on ants' collective problem solving ability. This global stochastic search method is inspired by the ability of a colony of ants to identify the shortest route between the nest and a food source, without using visual cues.

The operation mode of EAS algorithm is as follows: the artificial ants of the colony move, concurrently and asynchronously, through adjacent states of a problem, which can be represented in the form of a weighted graph. This movement is made according to a transition rule, called *random proportional* rule, through a stochastic mechanism. When ant $k$ is in node $i$ and has so far constructed the partial solution $s^p$, the probability of going to node $j$ is given by:

$$p_{ij}^{k} = \begin{cases} \dfrac{\tau_{ij}^{\alpha} + n_{ij}^{\beta}}{\sum\limits_{c_{il} \in N(s^p)} \tau_{il}^{\alpha} + n_{il}^{\beta}}, \text{ if } c_{ij} \in N(s^p) \\ 0 \qquad\qquad , \text{ otherwise} \end{cases} \quad (7)$$

where $N(s^p)$ is the set of feasible nodes when being in node $i$, i.e. edges $(i,l)$ where $l$ is the node not yet visited by the ant $k$. The parameters $\alpha$ and $\beta$ control the relative importance of the pheromone versus the heuristic information value $\eta_{ij}$, given by:

$$n_{ij} = \frac{1}{d_{ij}} \quad (8)$$

where $d_{ij}$ is the weight of each edge.

Individual ants contribute their own knowledge to other ants in the colony by depositing pheromones, which act as chemical "markers" along the paths they traverse. Through indirect communication with other ants via foraging behavior, a colony of ants can establish the shortest path between the nest and the food source over time with a positive feedback loop known as stigmergy. As individual ants traverse a path, pheromones are deposited along the trail, altering the overall pheromone density. More trips can be made along shorter paths and the resulting increase in pheromone density attracts other ants to these paths. The main characteristic of the EAS technique is that (at each iteration) the pheromone values are updated by all the $k$ ants that have built a solution in the iteration itself. The pheromone $\tau_{ij}$, associated with the edge joining nodes $i$ and $j$, is updated as follows [3]:

$$\tau_{ij} = (1-\rho)\cdot\tau_{ij} + \sum_{m=1}^{k}\Delta\tau_{ij}^{k} + \varepsilon\cdot\tau_{ij}^{elite} \tag{9}$$

where $\rho \in (0,1]$ is the evaporation rate, $k$ is the number of ants, $\varepsilon$ is the number of elitist ants, and $\Delta\tau_{ij}^{k}$ is the quantity of pheromone laid on edge (i, j) by ant k:

$$\Delta\tau_{ij}^{k} = \begin{cases} \dfrac{Q}{L_k}, \text{ if ant } k \text{ used edge } (i,j) \text{ in its tour} \\ 0 \quad \text{, otherwise} \end{cases} \tag{10}$$

where $Q$ is a constant for pheromone update, and $L_k$ is the length (or the weight of the edge) of the tour constructed by ant $k$. Furthermore, shorter paths will tend to have higher pheromone densities than longer paths since pheromone density decreases over time due to evaporation [3]. This shortest path represents the global optimal solution and all the possible paths represent the feasible region of the problem.

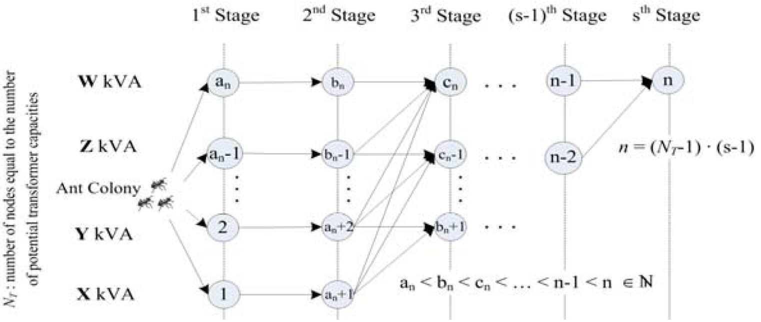
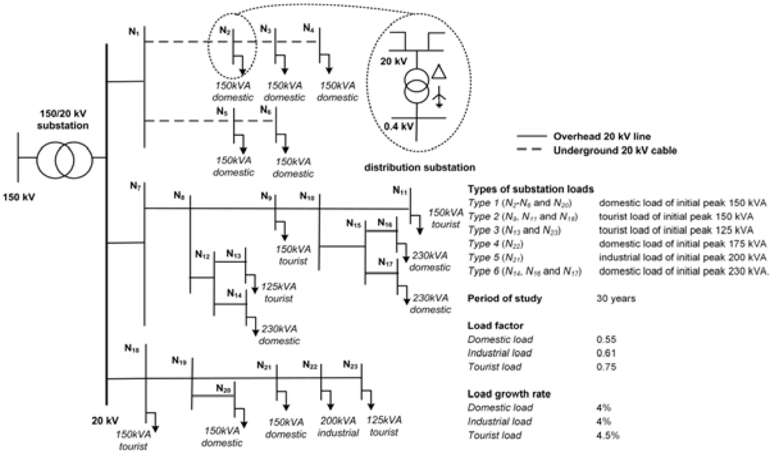


**Fig. 3.** The directed graph used for the OTS problem

### 5.2 OTS Implementation Using the EAS Algorithm

In this work our interest lies in finding the optimum choice of distribution transformers capacity sizing, so as to meet the load demand for all the years of the study period. To achieve so, a graph shown in Fig. 3 is constructed, representing the sizing paths.

The graph has *s* stages and each stage indicates a time period (in years) the limits of which are defined by the need to replace one of the considered transformer sizes due to violation of its thermal loading limits. Therefore, stage *s* has one node less in comparison with stage (*s-1*), stage (*s-2*) has one node less in comparison with stage (*s-1*), etc. The first stage indicates the beginning of the study, comprising number of nodes equal to the number of potential transformer capacities $N_T$, while *s* represents the end of the study period (consisting of the largest necessary rated capacity able to serve the load at the final year of the study). Symbols *X*, *Y*, *Z*, *W* refer to the different rated powers ( $X < Y < Z < W$ ). Furthermore, the arcs between the nodes are directed from the previous stage to the next one (backward movement is not allowed) since



| Types of substation loads | |
|---|---|
| Type 1 ($N_2$-$N_6$ and $N_{20}$) | domestic load of initial peak 150 kVA |
| Type 2 ($N_9$, $N_{11}$ and $N_{18}$) | tourist load of initial peak 150 kVA |
| Type 3 ($N_{13}$ and $N_{23}$) | tourist load of initial peak 125 kVA |
| Type 4 ($N_{22}$) | domestic load of initial peak 175 kVA |
| Type 5 ($N_{21}$) | industrial load of initial peak 200 kVA |
| Type 6 ($N_{14}$, $N_{16}$ and $N_{17}$) | domestic load of initial peak 230 kVA. |
| **Period of study** | 30 years |
| **Load factor** | |
| Domestic load | 0.55 |
| Industrial load | 0.61 |
| Tourist load | 0.75 |
| **Load growth rate** | |
| Domestic load | 4% |
| Industrial load | 4% |
| Tourist load | 4.5% |

**Fig. 4.** Single line diagram of the examined distribution network

each stage represents a forward step in the time of the study. Nodes 1 to $a_n$ (Fig. 3) are designated as the source nodes corresponding to each potential transformer size and node $n$ is designated as the destination node (Fig. 3). The objective of the colony agents is to find the least-cost path between nodes that belong to 1st stage and node $n$.

## 6 Results and Discussion

The proposed method is applied for the optimal choice of the transformer sizes of a practical distribution network. Fig. 4 illustrates the single line diagram of the examined distribution network. The substation type of loads and their initial peak value (at the first year of the study period) are indicated on the diagram (Fig. 4).

Six transformer ratings are considered, namely 160, 250, 300, 400, 500 and 630 kVA. Table 1 lists their main technical characteristics and bid price. The thermal calculations illustrated in Fig. 2 were repeated for the six transformers and each year of the study, resulting to the time periods of Table 1. The periods derived in Table 1 were used to define the stages of the graph of Fig. 3. In order to define the weight of each arc in the graph of Fig. 3, the energy loss cost calculation of each transformer for the studied period was based on the annual energy loss cost calculation described in Section 4. The energy loss cost corresponding to the transition from node $i$ ($p$-th year of the study period) to $j$ ($q$-th year of the study period), is computed by:

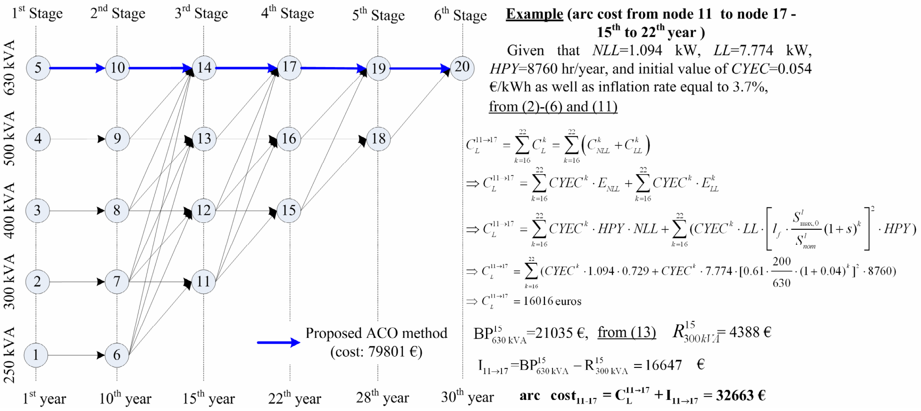
$$C_L^{i\to j} = \sum_{k=p+1}^{q} C_L^k \, . \tag{11}$$

When the transition from node $i$ to $j$ corresponds to transformer size upgrade from $S_i$ to $S_j$, the installation cost $I_{i\to j}$ derives from:

$$I_{i\to j} = BP_{S_j}^p - R_{S_i}^p \qquad (12)\,, \qquad R_{S_i}^p = BP_{S_i} \cdot \left[ \frac{(1+r)^N - (1+r)^m}{(1+r)^N - 1} \right] \qquad (13)$$

where $BP_{S_j}^p$ is the present value of the bid price of the transformer to be installed at the $p$-th year of the study period, $R_{S_i}^p$ is the remaining value of the uninstalled transformer, $BP_{S_i}$ is the transformer $S_i$ bid price, $r$ is the inflation rate, $N$ are the years of the transformer lifetime and $m$ are the years that the transformer was under service.

Table 2 lists the cost of four arcs of Fig. 5, calculated according to Section 4 and 6.



**Example (arc cost from node 11 to node 17 - 15th to 22th year )**

Given that $NLL$=1.094 kW, $LL$=7.774 kW, $HPY$=8760 hr/year, and initial value of $CYEC$=0.054 €/kWh as well as inflation rate equal to 3.7%, from (2)-(6) and (11)

$$C_L^{11\to17} = \sum_{k=16}^{22} C_L^k = \sum_{k=16}^{22}\left(C_{NLL}^k + C_{LL}^k\right)$$

$$\Rightarrow C_L^{11\to17} = \sum_{k=16}^{22} CYEC^k \cdot E_{NLL} + \sum_{k=16}^{22} CYEC^k \cdot E_{LL}^k$$

$$\Rightarrow C_L^{11\to17} = \sum_{k=16}^{22} CYEC^k \cdot HPY \cdot NLL + \sum_{k=16}^{22} (CYEC^k \cdot LL \cdot \left[ l_f \cdot \frac{S_{\max,0}^l}{S_{nom}^l}(1+s)^k \right]^2 \cdot HPY)$$

$$\Rightarrow C_L^{11\to17} = \sum_{k=16}^{22} (CYEC^k \cdot 1.094 \cdot 0.729 + CYEC^k \cdot 7.774 \cdot [0.61 \cdot \frac{200}{630} \cdot (1+0.04)^k]^2 \cdot 8760)$$

$$\Rightarrow C_L^{11\to17} = 16016 \text{ euros}$$

$\mathrm{BP}_{630\,\mathrm{kVA}}^{15}$=21035 €, from (13) $R_{300kVA}^{15}$= 4388 €

$\mathrm{I}_{11\to17} = \mathrm{BP}_{630\,\mathrm{kVA}}^{15} - \mathrm{R}_{300\,\mathrm{kVA}}^{15} = 16647$ €

**arc cost$_{11\text{-}17}$ = $\mathbf{C_L^{11\to17} + I_{11\to17}}$ = 32663 €**

**Fig. 5.** The directed graph used by the proposed ACO method for the Type 5 substation load

**Table 1.** Technical Parameters and Thermal Withstand of the Transformers Used in the Solution of the OTS Problem

| *Transformer technical parameters* | | | | *Transformer thermal withstand (yr) per type of substation load* | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| *Size (kVA)* | *Bid price (€)* | *NLL (kW)* | *LL (kW)* | *Type 1* | *Type 2* | *Type 3* | *Type 4* | *Type 5* | *Type 6* |
| 160 | 5275 | 0.454 | 2.544 | 5 | 4 | 8 | 3 | 0 | 0 |
| 250 | 6853 | 0.702 | 3.672 | 16 | 14 | 18 | 14 | 10 | 7 |
| 300 | 6932 | 0.738 | 4.186 | 21 | 18 | 23 | 19 | 15 | 12 |
| 400 | 9203 | 0.991 | 4.684 | 28 | 25 | 28 | 26 | 22 | 19 |
| 500 | 10296 | 1.061 | 5.771 | 30 | 30 | 30 | 30 | 28 | 24 |
| 630 | 12197 | 1.094 | 7.774 | - | - | - | - | 30 | 30 |

**Table 2.** Cost of indicative arcs in the graph of Fig. 5

| *Arc* | *Cost of the arc* | *Value (€)* | *Arc* | *Cost of the arc* | *Value (€)* |
|---|---|---|---|---|---|
| $2\to7$ | $C_L^{2\to7} + BP_{300}^{1st-year}$ | 17972 | $17\to19$ | $C_L^{17\to19}$ | 23601 |
| $11\to17$ | $C_L^{11\to17} + I_{11\to17}$ | 32663 | $19\to20$ | $C_L^{19\to20}$ | 11185 |

Fig. 5 illustrates the graph used by the proposed ACO method for the solution of the OTS problem for Type 5 substation load of Fig. 4 using the transformers of Table 1. We tested several values for each parameter, i.e. $\alpha\in\{0,0.5,1,2,5\}$, $\beta\in\{0,0.5,1,2,5\}$, $\rho\in\{0.1,0.3,0.5,0.7,1\}$. Table 3 includes the data of the optimal sizing strategies for each type of substation of Fig. 4 and Table 1. The optimal solutions of Table 3 were obtained using $k$=20, $\alpha$ =2, $\beta$ =0.5, $\rho$ =0.5, $Q$=2.7, max iterations=2000. According to Fig. 5, the optimal sizing path yielded by the proposed method corresponds to installation of the largest rated capacity that can serve the expected load at the end of the study period. The same optimal path is selected for the rest of types of the substation loads of the considered network, corresponding to the costs listed in Table 3. This is due to the fact that transformers with rated power significantly larger than the served load operate under low load current, thus consuming less annual energy losses (and consequently having less annual energy loss cost), in comparison with transformers of rated power close to the served load.

**Table 3.** Results of the proposed method for the OTS problem of Fig. 4

| ***Transformer cost*** | ***Substation load types*** | | | | | | |
|---|---|---|---|---|---|---|---|
| | ***Type 1*** | ***Type 2*** | ***Type 3*** | ***Type 4*** | ***Type 5*** | ***Type 6*** | ***Total*** |
| Installation cost (€) | 10677 | 10296 | 10677 | 10677 | 12648 | 12648 | 177573 |
| Energy loss cost (€) | 48529 | 61195 | 66372 | 56025 | 67153 | 70047 | 940822 |
| Total cost (€) | 59206 | 71491 | 77049 | 66702 | 79801 | 82695 | 1118395 |

## 7 Conclusions

In this paper, an EAS algorithm is proposed for the solution of the OTS planning problem by minimizing the overall transformer cost (i.e., the sum of the transformer purchasing cost plus the transformer energy loss cost) over the planning period, while satisfying all the problem constraints (i.e., the load to be served and the transformer thermal loading limit). The method is applied for the selection of the optimal size of the distribution transformers in a real network, comprising 16 distribution substations to serve a load over a period of 30 years. The application results show that the proposed EAS algorithm is very efficient because it always converges to the global optimum solution of the OTS problem.

**Acknowledgments.** This paper is part of the 03ED45 research project, implemented within the framework of the "Reinforcement Programme of Human Research Manpower" (PENED) and co-financed by National and Community Funds (25% from the Greek Ministry of Development-General Secretariat of Research and Technology and 75% from E.U.-European Social Fund).

## References

1. Chen, C.-S., Wu, T.-H.: Optimal Distribution Transformer Sizing by Dynamic Programming. Electrical Power & Energy Systems 20, 161–167 (1998)
2. Jovanovic, D.: Planning of Optimal Location and Sizes of Distribution Transformers using Integer Programming. Electrical Power & Energy Systems 25, 717–723 (2003)
3. Dorigo, M., Stützle, T.: Ant Colony Optimization. MIT Press, Cambridge (2004)
4. Stützle, T., Hoos, H.H.: MAX-MIN Ant System. Future Generation Computer Systems 16, 889–914 (2000)

5. Merkle, D., Middendorf, M., Schmeck, H.: Ant Colony Optimization for Resource-constrained Project Scheduling. IEEE Transactions on Evolutionary Computation 6, 333–346 (2002)
6. Leguizamon, G., Michalewicz, Z.: A New Version of Ant System for Subset Problems. In: Proc. of the 1999 Congress on Evolutionary Computation, vol. 2, pp. 1459–1464 (1999)
7. Blum, C.: Ant Colony Optimization: Introduction and Recent Trends. Physics of Life Reviews 2, 353–373 (2005)
8. Amoiralis, E.I., Tsili, M.A., Georgilakis, P.S., Kladas, A.G.: Ant Colony Solution to Optimal Transformer Sizing Problem. In: CD Proc. of EPQU (2007)
9. IEEE Guide for Loading Mineral-Oil-Immersed Transformers, IEEE Std C57.91 (2002)

# On the Evolutionary-Fuzzy Control of WIP in Manufacturing Systems

Nikos C. Tsourveloudis

Department of Production Engineering and Management
Technical University of Crete
73100 Chania, Crete, Greece
nikost@dpem.tuc.gr

**Abstract.** The effectiveness of optimized fuzzy controllers in the production scheduling has been demonstrated in the past through the extensive use of Evolutionary Algorithms (EA) for the Work-In-Process (WIP) reduction. The EA strategy tunes a set of distributed fuzzy control modules whose objective is to control the production rate in a way that satisfies the demand for final products, while reducing WIP within the production system. The EA identifies optimal design solutions in a given search space. How robust and generic is the controller that comes out of this process? This paper faces this question by testing the evolutionary tuned fuzzy controllers in demand conditions other than the ones used for their optimization. The evolutionary-fuzzy controllers are also compared to heuristically designed ones. Extensive simulations of production lines and networks show that the evolutionary-fuzzy strategy achieved a substantial reduction of WIP compared to the heuristic approach in all test cases.

**Keywords:** Manufacturing Systems, Work-In-Process, Fuzzy Control, Evolutionary Algorithms, Controller Design.

## 1 Introduction

As the manufacturing industry moves away from the mass production paradigm towards the agile manufacturing, the life cycle of products gets shorter while the need for a wide variety of them increases. Keeping large inventories in stock tends to be unattractive in today’s markets. The same holds for the unfinished parts throughout the manufacturing system, widely known as Work-In-Process (WIP), as it represents an already made expense with unknown profitability due to the rapidly changing demand. In a highly changing demand environment, the accumulated inventories are less desirable than ever.

The work-in-process inventory is measured by the number of unfinished parts in the buffers throughout the manufacturing system and it should stay as small as possible (for various reasons reported in [1], [2] and elsewhere).

Control policies aim in keeping WIP at low levels [3]. However, an exact optimal value of WIP cannot be determined in realistic manufacturing conditions. Therefore, the problem of WIP determination and control is amenable to an artificial intelligent treatment, as suggested in [4], [5] and recently in [6], [8].

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 18–25, 2008.

Fuzzy logic has been used in tandem to Evolutionary Algorithms (EA) so as to keep the WIP and cycle time as low as possible and at the same time to maintain high utilization [7], [9]. The objective in those works was to optimize the control policy in a way that satisfies the (random) demand for final products while keeping minimum WIP within the production system. During the evolution, the EA identifies those set of parameters for which the fuzzy controller has an optimal performance with respect to WIP minimization for several demand patterns.

The use of evolving genetic structures for the production scheduling problem, has recently gained a lot of acceptance in the automated and optimal design of fuzzy logic systems [10], [11]. However, a potential problem is that the evolutionary (or genetically) evolved fuzzy controllers might perform optimal only under the conditions involved in the evolution process. In this paper we examine the performance of evolutionary optimized controllers in contrast to heuristically designed fuzzy controllers. For comparisons purposes we test the controllers in conditions different from the ones they have been designed for. In this way, some useful insights regarding the design robustness of the evolutionary tuned fuzzy controllers may be drawn.

The rest of the paper is organized as follows. Section 2 describes the evolutionary fuzzy scheduling concept that is used for WIP minimization. Section 3 describes the comparison scenarios and presents experimental results for production lines and networks. Issues for discussion and remarks as well as suggestions for further development are presented in the last section.

## 2 Evolutionary-Fuzzy Scheduling

Traditionally, a production system is viewed as a network of machines and buffers. Items are received at each machine and wait for the next operation in a buffer with finite capacity. WIP may increase because of unanticipated events, like machine breakdowns and potential consequent propagation of these events. For example, a failed machine with operational neighbors forces to an inventory increase of the previous storage buffer. If the repair time is big enough, then the broken machine will either block the previous station or starve the next one. This “bottleneck” effect will propagate throughout the system.

Clearly, production scheduling of realistic manufacturing plants must satisfy multiple conflicting criteria and also cope with the dynamic nature of such environments. Fuzzy logic offers the mathematical framework that allows for simple knowledge representations of the production control/scheduling principles in terms of IF-THEN rules. The expert knowledge that describes the control objective (that is WIP reduction) can be summarized in the following statements [5], [8]:

*If the surplus level is satisfactory then try to prevent starving or blocking by increasing or decreasing the production rate accordingly,*

*else*

*If the surplus is not satisfactory that is either too low or too high then produce at maximum or zero rate respectively.*

In fuzzy logic controllers (FLCs), the control policy is described by linguistic IF-THEN rules similar to the above statements. The essential part of every fuzzy controller is the knowledge acquisition and the representation of the extracted knowledge with certain fuzzy sets/membership functions. Membership functions (MFs) represent the uncertainty modeled with fuzzy sets by establishing a connection between linguistic terms (such as low, negative, high etc) and precise numerical values of variables in the physical system. The correct choice of the MFs is by no means trivial but plays a crucial role in the success of an application. If the selection of the membership functions is not based on a systematic optimization procedure then the adopted fuzzy control strategy cannot guarantee minimum WIP level [9].

The evolutionary-fuzzy synergy attempts to minimize the empirical/expert design and create MFs that fit best to scheduling objectives [7], [9]. In this context, the design of the fuzzy controllers (distributed or supervisory) can be regarded as an optimization problem in which the set of possible MFs constitutes the search space. Evolutionary Algorithms (EAs) are seeking optimal or near optimal solutions in large and complex search spaces and therefore have been successfully applied to a variety of scheduling problems with broad applicability to manufacturing systems [10]. The objective is to optimize a performance measure which in the EAs context is called fitness function. In each generation, the fitness of every chromosome is first evaluated based on the performance of the production network system which is controlled through the membership functions represented in the chromosome. A specified percentage of the better fitted chromosomes are retained for the next generation. Then parents are selected repeatedly from the current generation of chromosomes, and new chromosomes are generated from these parents. One generation ends when the number of chromosomes for the next generation has reached the quota. This process is repeated for a pre-selected number of generations. The architecture of evolutionary-fuzzy WIP control scheme is presented in Fig. 1 and it is extensively discussed in [7] and [9].

The performance measure (fitness function) used in all previous treatments considers a known demand for products and the cumulative production of the system that produces these products. A typical fitness $F(x_i)$, of each individual $x_i$ is:

$$F(x_i)=\left[\sum_{j=1}^{N}(D(t_j)-PR(t_j))^2\right]^{-1}, \tag{1}$$

where, $t$ is the current simulation time, $T$ is the total simulation time and $D(t)$ is the overall demand and $PR(t)$ is the cumulative production of the system.

Assuming that the capacity of a production system is given, equation (1) shows that the evolved MFs are highly based (in terms of their support and shape) on the demand values. Some questions arise here: What if the actual demand is different (in both magnitude and changing pattern) than the one assumed in the evolution of the fuzzy controller? Is the evolved controller robust enough to absorb variations in demand? Or the original (without MF optimization) heuristic fuzzy performs better in unknown demands? Since there are no analytical solutions to those questions, in what follows we will examine and compare the performance of both evolutionary and heuristic fuzzy controllers through simulation, for a wide variety of test cases.

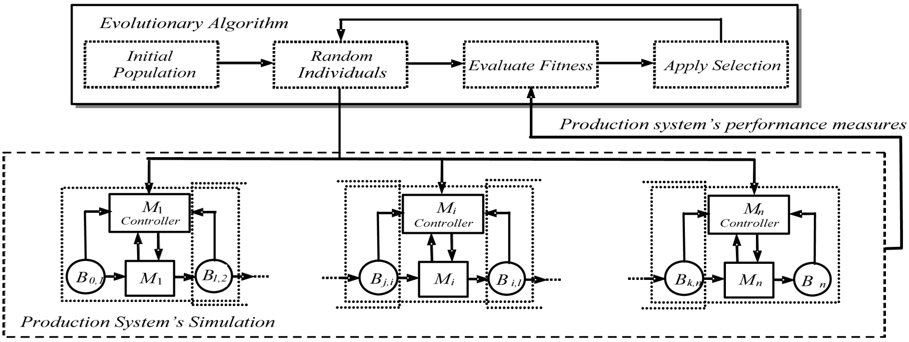


**Fig. 1.** Evolutionary-fuzzy control concept

## 3 Testing and Results

The evolutionary-fuzzy approaches suggested in [7], are tested and compared to the heuristic fuzzy approaches initially suggested in [5]. In the all simulations performed we assume that the machines fail randomly with a failure rate $p_i$. This rate is known and set before the simulation starts. Also, machines are repaired randomly with rate $rr_i$. The resources needed for repairs are assumed to be unlimited. The times between failures and repairs are exponentially distributed. All machines operate at known, but not necessarily equal rates. Each machine produces in a rate $r_i \leq \mu_i$, where $\mu_i$ is the maximum processing rate of machine $M_i$. We also assume that the flow of parts within the system is continuous.

The initial buffers are infinite sources of raw material and consequently the initial machines are never starved. The buffer levels at any time instant are given by:

$$b_{j,i}(t_{k+1}) = b_{j,i}(t_k) + [r_j(t_k) - r_i(t_k)](t_{k+1} - t_k), \tag{2}$$

where $t_k$, $t_{k+1}$, $r_i$ are the times when control actions (changes in processing rates) happen. The cumulative production of a machine $M_i$ is

$$PR_i(t_{k+1}) = PR_i(t_k) + r_i(t_k)(t_{k+1} - t_k). \tag{3}$$

In all simulations runs set-up and transportation times are negligible or included in the processing times. Buffers between adjacent machines $M_i$, $M_j$ assumed to have finite capacities.

Two common layouts of a production system are considered. A production line (Fig. 2a) and a production network (Fig. 2b). In Figure 2, circles represent buffers and the squares are machines. For simplicity both systems are assumed to produce one part type. Lines and networks producing multiple part types have been discussed in [5], [6] and it has been shown that have similar behavior to the single-part-type systems. The production systems of figure 2 are identical to the systems discussed

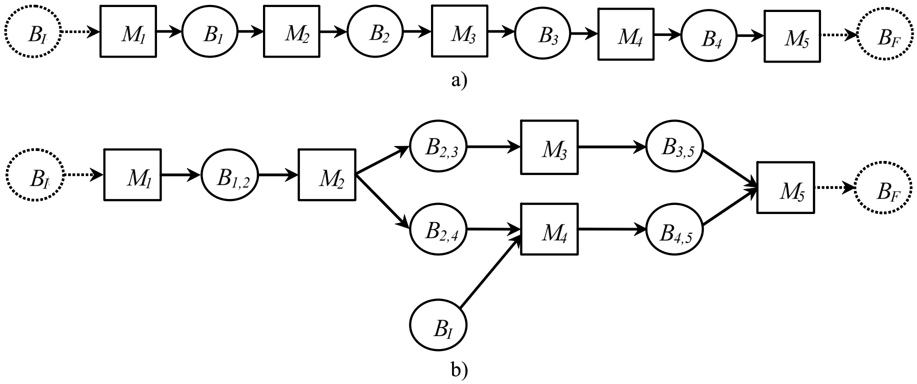


**Fig. 2.** The production systems used for controllers testing: a) Line, b) Network

in [5], [6], [7], [9]. This was selected on purpose so as to facilitate the comparisons with previous approaches. The main observation made in [6], [7] and [9] was that the evolutionary tuned fuzzy controllers achieved a substantial reduction of WIP in almost all test cases. This is expected since the controllers were evolved for known patterns of demand that is either constant or stochastic with certain mean values. In the test cases that follow, we keep unaltered the controllers' design but we scientifically change the demand patterns. In practice, and of course depending on the product, demand is the main uncertainty that comes from the outside of the production system.

### 3.1 Test Case 1: Production Lines

The production line under consideration (Fig. 2a) consists of five machines producing one product type. The failure and repair rates are equal for all machines. The repair rates are $rr_i$=0.5 and the failure rates are $p_i = 0.1$. The processing rates are also equal for all machines and are equal to $\mu_i = 2$ ($i$=1,...,5). All buffer capacities are equal to $BC_i = 10$.

In the evolution of the original fuzzy controllers for production lines, the demand was either considered constant (specified items per time unit) or stochastic (with known mean values and a small variation). Now the demand patterns are significantly changed, as can be observed in Figure 3. The value of $\overline{WIP}$ for both evolutionary (EFC: Evolutionary Fuzzy Controller) and heuristic (HFC: Heuristic Fuzzy Controller) is presented also in Figure 3. As can be seen the demand is far from being constant. For testing purposes, the demand shown in Figure 3 takes a random value between zero and 2.5 items every 20 time units. It has been observed that both controllers satisfy the demand and the same time achieve low WIP levels. But the evolutionary tuned is better than the heuristic one in the long run. This was the case in various tests with multiple changes in demand. As shown in Figure 4, for a more frequently changing demand, the evolutionary tuned controller is better in keeping WIP low than the heuristically designed controller.

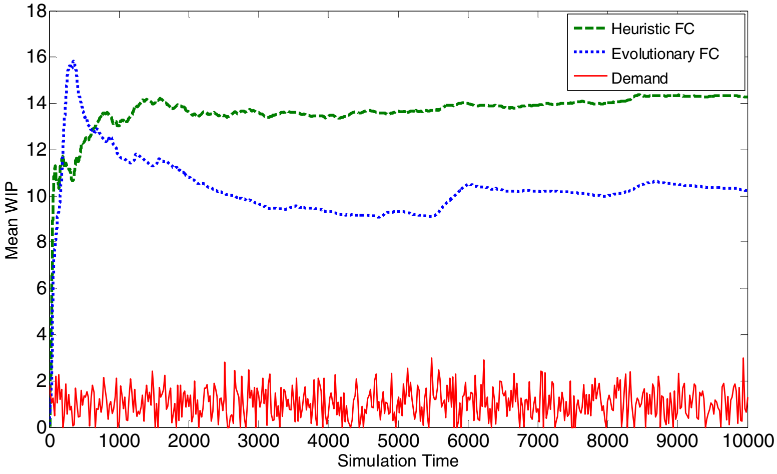


**Fig. 3.** Evolution of $\overline{WIP}$ in test case 1: Demand changes every 20 time units

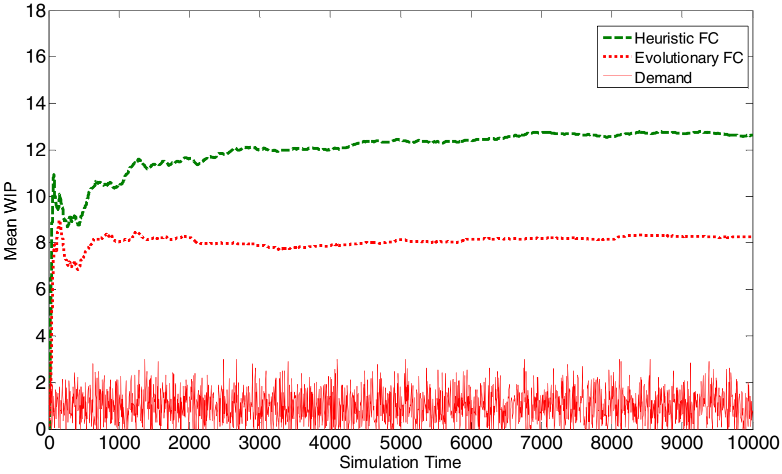


**Fig. 4.** Evolution of $\overline{WIP}$ in test case 1: Demand changes every 5 time units

### 3.2 Test Case 2: Production Networks

The production network (Fig. 2b) consists of five machines also producing one part type. The failure and repair rates of all machines are equal. The repair rates are $rr_i$= 0.5 and the failure rates are $p_i = 0.1$. The processing rates are also equal for all machines and are equal to $\mu_i = 5$ ($i$=1,..., 5). All buffer capacities are equal to $BC_i = 10$.

As expected (and may be seen in Fig. 5), the $\overline{WIP}$ levels in test case 2 (production network) are higher than in the test cases 1 (production line). Also in test case 2 the

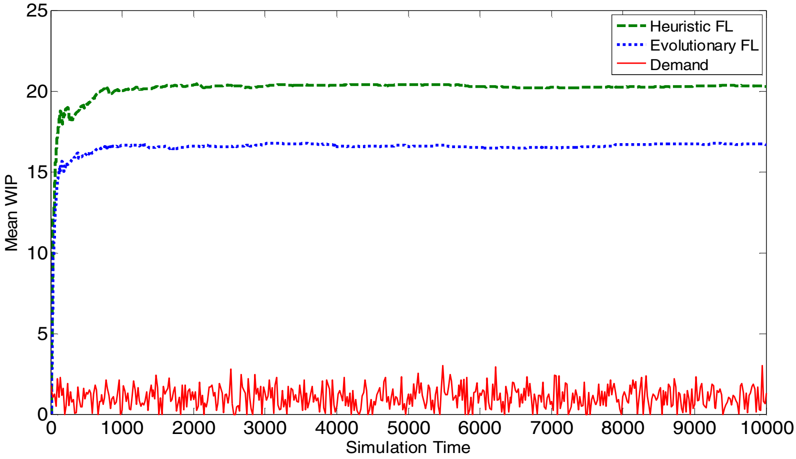


**Fig. 5.** Evolution of $\overline{WIP}$ in test case 2: Demand changes every 20 time units

evolutionary fuzzy controller gave less WIP than the heuristic fuzzy controller regardless of the demand changing frequency.

## 4 Observations and Concluding Remarks

A remarkable control ability of the WIP is shown in cases with a frequent demand change. This ability was observed regardless of production system's design complexity, as in both lines and networks the WIP is substantially reduced compared to the empirical selected fuzzy controllers.

It is known that WIP itself cannot represent adequately of production system's performance. One has to take into account also the accumulated orders backlog. It is also known that when demand is very high one may consider that service rate and thus backlog is more important than WIP. When demand can be easily satisfied and backlog is in low levels, a substantial reduction of WIP may be more important than a small increase in backlog. What we have seen so far is that with the aid of the evolutionary-fuzzy controllers the system's performance becomes more balanced in terms of mean WIP and backlog.

The heuristic fuzzy control approach cannot achieve the performance of the evolutionary-fuzzy. However, it is still better than previously reported "bang-bang" control approaches. Even when compared to the evolutionary-fuzzy approach it is much simpler in the design process as it steps on the human expertise/knowledge regarding the production system. In others words, one should very fast design, built and put to work a fuzzy controller with membership functions that represent the expert knowledge in contrast to the evolutionary-fuzzy system whose parameters are automatically set by the optimization procedure.

The evolutionary-fuzzy controllers are capable of maintaining low WIP levels for product demands other than the ones used during the optimization. Therefore, the evolutionary algorithms clearly represent a successful approach towards the optimization of robust scheduling approaches.

An interesting future extension of this work might be the use of EA strategies in more complex production systems such as multiple-part-type and/or reentrant systems.

**Acknowledgments.** This work was supported by the Hellenic General Secretariat for Research and Technology, grant E.P.A.N. M.4.3.6.1, 2004ΣE 01330013, Project code: 049-γ.

## References

1. Conway, R., Maxwell, W., McClain, J.O., Joseph Thomas, L.: The role of work-in-process inventory control: single-part-systems. Oper. Res. 36, 229–241 (1988)
2. Bai, S.X., Gershwin, S.B.: Scheduling manufacturing systems with work-in-process inventory control: multiple-part-type systems. Int. J. Prod. Res. 32, 365–386 (1994)
3. Gershwin, S.B.: Manufacturing Systems Engineering. Prentice Hall, New Jersey (1994)
4. Custodio, L., Sentieiro, J., Bispo, C.: Production planning and scheduling using a fuzzy decision system. IEEE Trans. Robot. Automat. 10, 160–168 (1994)
5. Tsourveloudis, N.C., Dretoulakis, E., Ioannidis, S.: Fuzzy work-in-process inventory control of unreliable manufacturing systems. Inf. Sci. 27, 69–83 (2000)
6. Ioannidis, S., Tsourveloudis, N.C., Valavanis, K.P.: Fuzzy Supervisory Control of Manufacturing Systems. IEEE Trans. Robot. Automat. 20, 379–389 (2004)
7. Tsourveloudis, N.C., Doitsidis, L., Ioannidis, S.: Work-in-Process Scheduling by Evolutionary Tuned Distributed Fuzzy Controllers. In: Proceedings of the IEEE International Conference on Robotics and Automation, Orlando, FL, USA, May 15-19 (2006)
8. Tsourveloudis, N.C., Ioannidis, S., Valavanis, K.P.: Fuzzy Surplus based Distributed Control of Manufacturing Systems. Advances in Production Engineering and Management 1, 5–12 (2006)
9. Tsourveloudis, N.C., Doitsidis, L., Ioannidis, S.: Work-In-Process Scheduling by Evolutionary Tuned Fuzzy Controllers. International Journal of Advanced Manufacturing Technology 34 (2007)
10. Tedford, J.D., Lowe, C.: Production scheduling using adaptable fuzzy logic with genetic algorithms. Int. J. Prod. Res. 41, 2681–2697 (2003)
11. Gordon, O., Herrera, F., Hoffmann, F., Luis, M.: Genetic Fuzzy Systems: Evolutionary Tuning and Learning of Fuzzy Knowledge Base. World Scientific Publishing Co. Pte. Ltd, U.K (2001)

# Numerical Study and Optimal Blade Design of a Centrifugal Pump by Evolutionary Algorithms

Vasilis Grapsas, Fotis Stamatelos, John Anagnostopoulos, and Dimitris Papantonis

School of Mechanical Engineering / Fluids Section
National Technical University of Athens
9 Heroon Polytechniou ave., Zografou, 15780 Athens, Greece

**Abstract.** A numerical methodology for hydrodynamic design in centrifugal pumps is developed and tested, considering the possibility to increase the hydraulic efficiency of the pump impeller by improving the blades shape. The simulation and analysis of the incompressible turbulent flow through the test impeller is performed with a commercial CFD code, and the numerical results are in agreement with the corresponding measurements in a laboratory pump impeller. A parametric study is carried out to examine the influence of some blade design parameters on the performance and the efficiency of the impeller, like the blade length, the inlet height, and the leading edge inclination. The values of the above parameters that maximize the hydraulic efficiency of the impeller are then derived by a multiparametric optimization methodology, using a stochastic evolutionary algorithm software. The optimal impeller design exhibits a remarkable efficiency increase compared to the initial impeller design.

**Keywords:** Centrifugal pump impeller, Numerical modeling, Design optimization, Evolutionary algorithms.

## 1 Introduction

Computational Fluid Dynamics (CFD) analysis is being lately increasingly applied in the design of centrifugal pumps, and nowadays the complex internal flows in water pump impellers can be predicted to speed up the pump design procedure [1-4]. On the other hand, the design of a component such as a radial pump impeller is a formidable challenge for designers. The main reason is that in order to be able to generate a large panel of blade geometries, a large set of geometrical parameters is needed [5]. A numerical analysis with such a number of parameters is time consuming, without guaranteed convergence to an acceptable solution. Therefore, a CFD code must be combined with a reliable optimization technique in order to reduce the computing cost and the design time. Various optimization techniques are becoming more and more standard for this purpose [6-7].

Examples of methods for optimal design of centrifugal pump impellers are today not so numerous. A design method for blade profiles of a centrifugal pump impeller based on a parallel optimization algorithm in combination with CFD was presented by

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 26–33, 2008.

Wahba [8]. The impeller's blades were represented with Bezier curves achieving satisfactory designs effectively. For the design optimization of a first stage pump impeller, Visser [9] used a blade angle distribution for the construction of the blades profile.

In the present paper the numerical results of the flow in a centrifugal pump with single circular arc blades are compared to the experimental measurements of a similar impeller constructed and operating in the Laboratory [10]. Then, the numerical code is used to perform parametric studies of the 3D blade geometry, and finally a numerical optimization of the blade design is carried out. This study is a first step towards a more elaborate parameterization of a centrifugal pump impeller, where several additional variables will be introduced in order to obtain twisted blade shapes and to describe the blade surfaces with Bezier polynomials. The developed methodology is based on the combination of a parametric code, which creates the 3D impeller geometry, a general optimization algorithm (EASY [5]), and a commercial CFD code (Fluent®). A similar methodology has been recently applied by the authors for the optimal design of a 2D impeller using home-made flow solver. [11-12].

## 2 Laboratory Data

The experimental apparatus is described in detail elsewhere [10], and only some technical characteristics are repeated here in brief. Figure 1a shows a cross-sectional view of the experimental test section, with a vertical rotating shaft, which was constructed and installed in the Laboratory of Hydraulic Turbomachinery, NTUA. The diffuser and the casing are made of plexiglass to enable direct observations of the flow. The so formed casing is axisymmetric and the generated flow field exhibits periodic symmetry. The measuring equipment is composed of two differential pressure transducers, one for the head of the pump (by measuring the pressure difference in points “1” and “2”, Fig. 1a), and the second for the flow rate (through the pressure difference across a calibrated orifice plate), a torsional torque meter installed at the shaft between the impeller and the motor, and a digital counter for the speed of rotation.

A radial flow, centrifugal pump impeller designed and constructed in the Lab [10] was tested (Fig. 1b). The inlet and outlet diameters are 70 mm and 190 mm, respectively, and the corresponding blade angles are 26$^o$ and 40$^o$. It has 9 blades and its nominal flow rate and head at the Best Efficiency Point (BEP) are 62,5 m$^3$/h and 47,5 m, respectively, at 3000 rpm.

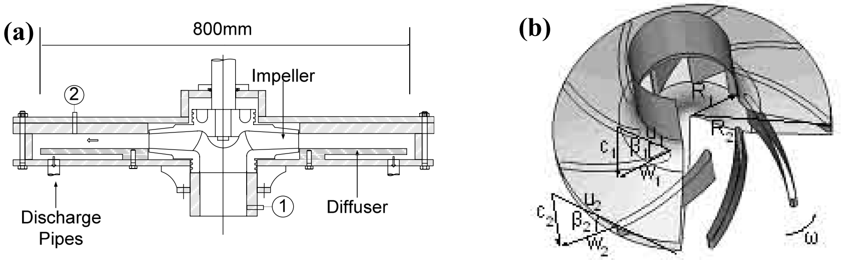


**Fig. 1.** The model pump: a) Cross sectional view; b) 3D sketch of the impeller

# 3 Numerical Method

## 3.1 Flow Solver

The flow and pressure field through the impeller results from the solution of the fully three-dimensional incompressible Navier-Stokes equations, while turbulence is simulated with the standard *k-ε* model. The simulation is performed with the commercial CFD code Fluent®, which has been widely used in the field of turbomachinery with reliable results [13, 14]. The pressure-velocity coupling is implemented through the SIMPLE algorithm, using second order upwind discretization for the convection terms and central differences for the diffusion terms.

Wall functions based on the logarithmic law are imposed at all solid walls, and periodic boundary conditions are used since the computational domain includes only one blade (Fig. 2a). No slip conditions are taken at all solid, stationary or rotating surfaces, while uniform inflow and free stream conditions are set at the flow inlet and outlet, respectively.

The computational grid is of structured type and it is generated with the Fluent preprocessor Gambit. The whole domain consists of a lot of subdomains or zones, in a way that the density and the quality of the cells in local flow field regions can be suitably controlled and handled depending on pressure gradients and velocities. The numerical grid used for the present calculations has a number of approximately 800.000 cells, which was found good enough as far as accuracy and computation cost are concerned. An indicative view of the mesh around the blade is shown in Fig. 2a, and a detail view at the leading edge in Fig. 2b. For stability reasons, the algorithm starts from low rotational speed and gradually reaches the nominal speed of 3000 rpm.

## 3.2 Impeller Head and Power Calculations

Using a converged flow field, the energy gained by the fluid through the impeller, *H*, is computed from the total energy of the fluid at the impeller inlet and exit periphery, as shown in Fig. 1b:

$$H = H_2 - H_1 = \frac{1}{Q_u} \cdot \left[ \int_0^{2\pi} \left( \frac{p_2}{\rho g} + \frac{c_2^2}{2g} \right) \cdot c_{r2} \cdot b_2 \cdot r_2 \cdot d\theta - \int_0^{2\pi} \left( \frac{p_1}{\rho g} + \frac{c_1^2}{2g} \right) \cdot c_{r1} \cdot b_1 \cdot r_1 \cdot d\theta \right] \quad (1)$$

where *c* is the absolute fluid velocity $c_r$ its radial component, *p* the pressure, *b* the blade height, $Q_u$ the flow rate through the impeller and *g* the gravity acceleration, while the subscripts 1 and 2 denote impeller inlet and exit conditions, respectively (Fig. 1b). The impeller head $H_u$ can be calculated from the torque $M_u$ developed on the blades:

$$N_u = \rho \cdot g \cdot Q_u \cdot H_u = \omega \cdot M_u = \omega \cdot \int_{r_1}^{r_2} \left[ \left( \vec{r} \times \vec{n} \right) \cdot p + \left( \vec{r} \times \vec{\tau}_w \right) \cdot cot\,\beta \right] \cdot b \cdot dr \quad (2)$$

where $\vec{n}$ the unit vector normal to the blade surface, $\vec{\tau}_w$ the wall shear stress, $\beta$ the blade angle and *b* the blade height (or the impeller width). The hydraulic efficiency of the impeller can be finally obtained from its definition:

$$\eta_h = H / H_u \quad (3)$$

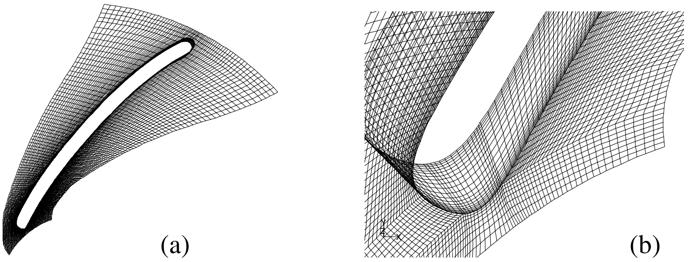


**Fig. 2.** a) Computational mesh, and b) detailed view at the leading edge

## 4 Validation of the Simulation Method

At first the numerical model is applied to calculate the flow field developed in a standard impeller at constant rotational speed of 3000 rpm, the blades of which have a constant height and their mean camber line follows a single circular arc curvature. The nominal flow rate and the other impeller technical specifications comply with the corresponding laboratory model pump. Some indicative numerical results are drawn in Fig. 3.

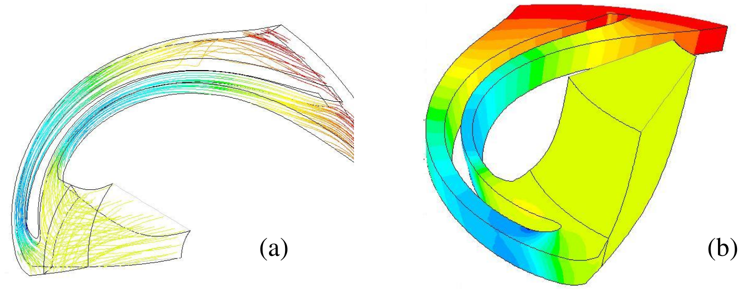


**Fig. 3.** Indicative numerical results: a) flow streamlines, and b) Pressure contours

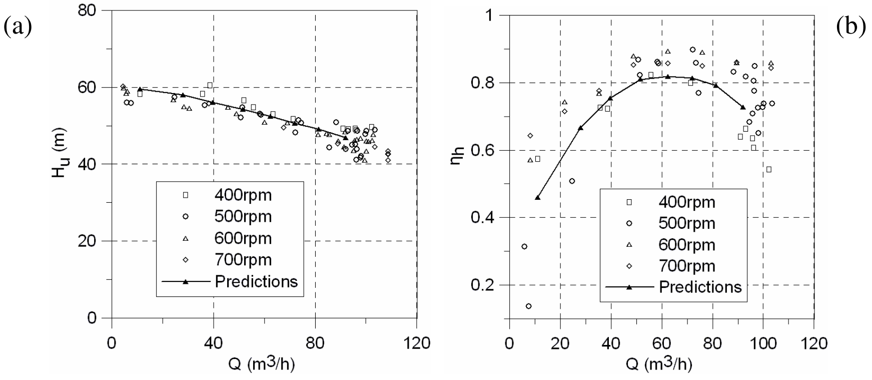


**Fig. 4.** Measured and predicted impeller values: a) Head; b) Hydraulic efficiency

On the other hand, the experiments were carried out at four different rotation speeds: 400, 500, 600 and 700 rpm. The obtained quantities, net fluid head *H*, impeller power $N_u$ and hydraulic efficiency $\eta_h$ were then transformed to 3000 rpm, through the affinity laws. The measured and computed characteristic curves *H-Q,* and $\eta_h$-*Q* of the impeller are compared in Fig. 4, where the agreement is quite encouraging. The predicted lower efficiency around and above the BEP are due to the increased losses at the impeller entrance, where the simulated impeller has smaller width and sharper corners than the model.

# 5 Results

## 5.1 Parametric Studies

The 3D blade geometry is described here by three design parameters: The blade length, the blade inlet height, and the inclination of the blade leading edge. The effect of each of these on the impeller performance is investigated first, using the flow analysis software. The blade length is a basic design parameter, which affects also the passage width between two consecutive blades. The blade length can be varied for constant inlet and outlet blade diameters and angles, and it can be defined through the wrap angle $\theta_w$, or the overlap angle $\theta$ of the blades (Fig. 5).

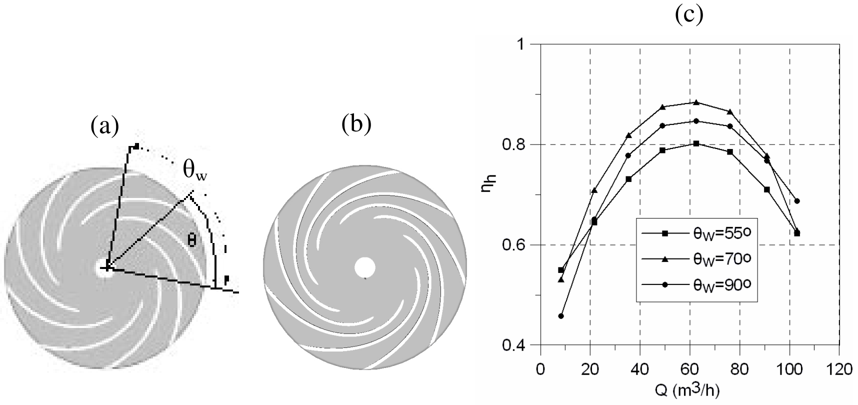


**Fig. 5.** Blade length definition and variation: a) $\theta_w$=55°; b) $\theta_w$=90°; c) Comparative results

The blade profile geometry is generated by the point by point method [12], in which the blade angle variation between given inlet and outlet angles is described by a second order polynomial, and the remaining free variable determines here the wrap angle, or the blade length.

Three impellers having blades of different wrap angles ($\theta_w$=55°, 70° and 90°) were simulated and the computed characteristic curves for the hydraulic efficiency are drawn in Fig. 5c. The wrap angle seems to exert a great influence to the efficiency, and the results reveal the existence of an optimum wrap angle, around 70°, for which the impeller efficiency is maximized.

The correlation of hydraulic efficiency with the blade inlet height (or the impeller inlet width) is studied next. The standard blade geometry has a constant height $b_1$=$b_2$ (Fig. 6a), whereas in an alternative design the blade height follows a linear distribution, starting from $b_1$=2$b_2$ at the leading edge (Fig. 6b).

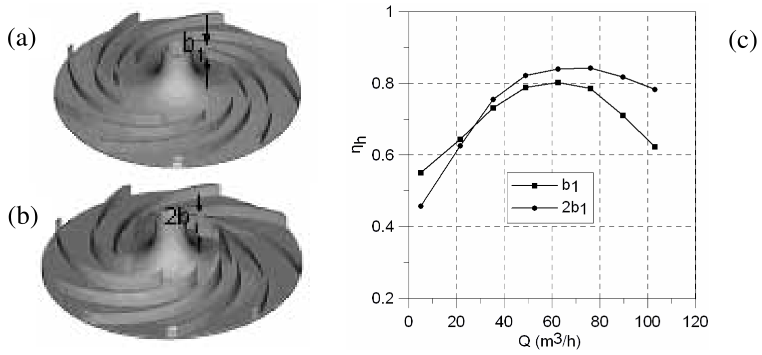


**Fig. 6.** Blade inlet height: a) constant; b) linearly varied; c) Comparative results

The head-flow curves of these two impellers do not present substantial differences around the BEP, whereas a remarkable increase in hydraulic efficiency can be observed in Fig. 6c for the impeller having wider inlet. Consequently, the optimization of this parameter should be also considered in the design process.

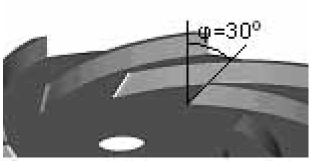


**Fig. 7.** Blades with inclined leading edge

For the non-twisted geometry of the blades considered here, the use of an inclined leading edge, as shown in Fig. 7, results in variable inlet radius along the blade height, and hence in variation of the relative fluid velocity angle at the impeller inlet. The latter is expected to cause increased mechanical losses at the inlet, and this was verified numerically: The hydraulic efficiency of the impeller having a $\varphi$=30° inclined blade edge is about 10% reduced compared to $\varphi$=0, while the BEP is displaced to higher flow rates. Nevertheless, this parameter is also included in the design variables in order to test the reliability of the subsequent optimization process and the ability of the optimizer to find or approach the optimum value $\varphi = 0$.

## 5.2 Design Optimization

The above three geometric parameters constitute the free design variables, which can be modified within limits, in order to find the optimum blade shape that maximizes the hydraulic efficiency of the impeller at the BEP. The Evolutionary Algorithms SYstem (EASY) optimization software used in this study is recently developed and brought to market by the Laboratory of Thermal Turbomachinery, NTUA [5]. It makes use of stochastic optimization methods and looks automatically for the set of the design variables that maximizes the cost function, using populations of candidate solutions instead of a single solution. The passage from a population set to the next one mimics the biological evolution of species generations.

As shown in Fig. 8a, the optimizer converges after about 100 completed solutions of the flow field to the optimum impeller, which has the following characteristics: blade wrap angle $\theta_w$=74°, inlet height $b_1$=1,2$b_2$, and leading edge almost vertical (φ=2,1°), as expected. The relatively great number of iterations for convergence is mainly due to the use of wide variation limits for the design variables (50° – 90° for $\theta_w$, 1 – 3$b_1$ for the inlet height, and 0 – 50° for the leading edge slope).

The improved performance of the optimal design is shown in Fig. 8b, in comparison with the model impeller predictions and measurements. The hydraulic efficiency of the improved impeller is more than 10 percentage units higher than that of the original impeller around the BEP, and remains higher in the entire flow rate range (Fig. 8b). It is also clearly above the measured values of the model impeller, as well as of all the rest corresponding curves of the impellers examined during the parametric study (Figs. 5c and 6c). Therefore, the reliability of the developed multiparametric optimization methodology is verified.

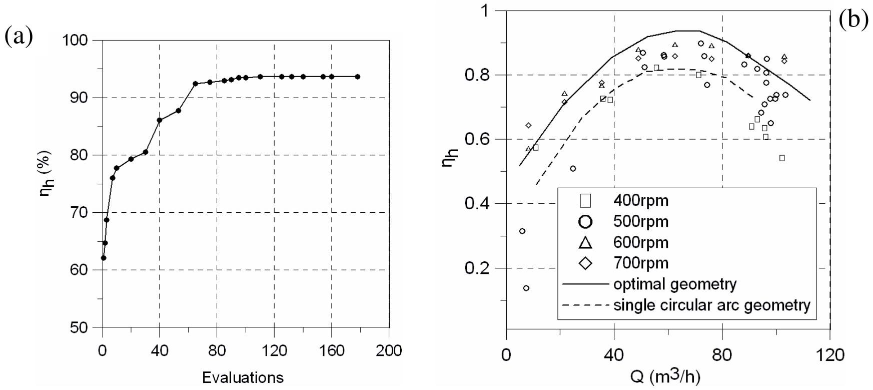


**Fig. 8.** Original and optimum impeller efficiency

## 6 Conclusions

A commercial flow analysis software is used to conduct parametric studies of the effect of some geometric parameters of a centrifugal pump impeller, and the results revealed that their modifications can have a significant impact on its performance.

In order to maximize the hydraulic efficiency of the impeller, a design optimization methodology is developed and applied using a stochastic optimization software with evolutionary algorithms. The results of a 3-parametric optimization study showed that the derivation and use of optimal values for these parameters can increase substantially the efficiency.

Additional free design variables can be easily handled by the present methodology, and their incorporation is expected to improve even more the impeller design. Moreover, multi-objective optimization is also possible and it is planned for the future, in order to design an impeller that combines the improved performance and efficiency with a cost-effective construction.

## References

1. Hornsby, C.: CFD – Driving pump design forward. World Pumps, pp. 18–22 (2002)
2. Zhang, M.J., Gu, C.G., Miao, Y.M.: Numerical study of the internal flow field of a centrifugal impeller. ASME Paper 94-GT-357 (1994)
3. Majidi, K.: Numerical study of unsteady flow in a centrifugal pump. ASME Journal of Turbomachinery 127, 363–371 (2005)
4. Anagnostopoulos, J.: CFD analysis and design effects in a radial pump impeller. WSEAS Trans. On Fluid Mechanics 7(1), 763–770 (2006)
5. Giannakoglou, K.C.: Design of optimal aerodynamic shapes using stochastic optimization methods and computational intelligence. Progress in Aerospace Sciences 38, 43–76 (2002)
6. Cao, S., Peng, G., Yu, Z.: Hydrodynamic design of rotodynamic pump impeller for multiphase pumping by combined approach of inverse design and CFD analysis. ASME Transactions, Journal of Fluids Engineering 127, 330–338 (2005)
7. Asuaje, M., Bakir, F., Kouidri, S., Rey, R.: Inverse design method for centrifugal impellers and comparison with numerical simulation tools. International Journal for Computational Fluid Dynamics 18(2), 101–110 (2004)
8. Wahba, W., Tourlidakis, A.: A Parallel Genetic Algorithm Applied to the Design of Blade Profiles for Centrifugal Pump Impellers. American Institute of Aeronautics & Astronautics (1998)
9. Visser, F., Dijkers, R., Woerd, J.: Numerical flow-field analysis and design optimization of a high-energy first-stage centrifugal pump impeller. Computing and Visualization in Science 3, 103–108 (2000)
10. Grapsas, V., Mentzos, M., Anagnostopoulos, J., Filios, A., Margaris, D., Papantonis, D.: Experimental and Computational Study of Radial Flow Pump Impeller. In: Proceedings 2nd IC-SCCE, Athens, Greece (2006)
11. Grapsas, V., Anagnostopoulos, J., Papantonis, D.: Parametric Study and Design Optimization of Radial Flow Pump Impeller by Surface Parameterization. In: Proceedings 2nd IC-SCCE, Athens, Greece (2006)
12. Grapsas, V., Anagnostopoulos, J., Papantonis, D.: Hydrodynamic Design of Radial Flow Pump Impeller by Surface Parameterization. In: Proceedings 1st IC-EpsMsO, Athens, Greece (2005)
13. Sun, J., Tsukamoto, H.: Off-design performance prediction for diffuser pump. Journal of Power and Energy 215, 191–201 (2001)
14. Gonzalez, J., Fernandez, J., Blanco, E., Santlaria, C.: Numerical Simulation of the dynamic effects due to impeller-volute interaction in a centrifugal pump. ASME Journal of Fluids Engineering 125, 348–355 (2002)

# Groundwater Numerical Modeling and Environmental Design Using Artificial Neural Networks and Differential Evolution

Ioannis K. Nikolos[1], Maria Stergiadi[2], Maria P. Papadopoulou[2],
and George P. Karatzas[2]

[1] Department of Production Engineering and Management, Technical University of Crete,
University Campus, Kounoupidiana, 73100, Chania, Greece
[2] Department of Environmental Engineering, Technical University of Crete, University
Campus, Kounoupidiana, 73100, Chania, Greece
jnikolo@dpem.tuc.gr, karatzas@mred.tuc.gr

**Abstract.** A Differential Evolution (DE) algorithm is combined with an Artificial Neural Network (ANN) to examine different operational strategies for the productive pumping wells located in the Northern part of Rhodes Island in Greece. The objective is to maximize the pumping rate without violating the environmental constraints associated with the water table drawdown at critical locations. The hydraulic head field is simulated using a groundwater flow simulator that solves numerically a system of partial differential equations. Successive calls to the simulator are used to provide the training data to the ANN. Then the ANN is used as an approximation model to the simulator, successively called by the DE algorithm to evaluate candidate solutions. The adopted procedure provides the ability to test different scenarios, concerning the optimization constraints, without retraining of the ANN, which significantly reduces the computational cost of the procedure.

**Keywords:** Artificial Neural Networks, Differential Evolution, groundwater management.

## 1 Introduction

Keeping the coastal water resources in good quality is of primary importance for both the human communities and the environment; the sudden lowering of the water table, due to continuous and unrestrained pumping activity, besides the effects on the water quality, may also affect the hydraulic connection between the surface and subsurface water resources [1]. During the recent years ANNs have been introduced in the field of water resources modeling and management. The research cited (and the work contained in this paper) pertain only to the Multi Layer Perceptron type of ANNs. Rogers and Dowla [2] introduced ANNs in a two-step approach to non-linear groundwater management problems. The optimal solutions were obtained by first training an ANN to predict the flow and transport outcomes of various pumping strategies. Then, the

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 34–41, 2008.

ANN searches for the optimal solution through the various pumping schemes. Rogers et al. [3] also applied a combination of an ANN with a Genetic Algorithm (GA) to a field-scale groundwater-remediation design problem. Morshed and Kaluarachchi [4] used an ANN to simulate the responses of the physical system, using a limited number of groundwater flow and contaminant transport simulation to considerably reduce the computational time. An ANN based multi-objective optimization model was introduced by Wen and Lee [5] to predict decision maker's preferences for water quality management in river basins. A combined GA-ANN methodology was presented by Aly and Peralta [6] to account for uncertainty in hydraulic conductivity in a field-scale problem. The GA is used to compute the optimal pumping strategy while the ANN is used to estimate the concentration response surface within the GA. Rao et al. [7] developed a subsurface water management model applied to Deltaic regions by interfacing Simulated Annealing algorithm with an existing sharp interface flow model. Arndt et al. [8] proposed an approach that utilizes ANN computed predictions to approximate the results of a computationally expensive finite-element simulation model. Up to 60% time reduction was obtained using the ANN prediction solution, compared to the simulation-based solution. Yan and Minsker [9] introduced a dynamic modeling approach in which ANNs were adaptively and automatically trained directly within a GA to replace the time-consuming water resources simulation models.

In this work a Differential Evolution (DE) algorithm is combined with an ANN to provide a tool for the fast testing of different optimal operational strategies for pumping wells; the procedure is implemented for the productive wells located in the Northern part of Rhodes Island in Greece. The objective is to maximize the pumping rate of the productive wells, in order to cover the water demand, without violating the environmental constraints associated with the water table drawdown at critical locations. The hydraulic head field is simulated using a groundwater flow simulator that numerically solves a system of partial differential equations. Successive calls to the simulator, for different pumping rates, are initially used to provide sufficient training data to the ANN. Then the ANN is used as an approximation model to the simulator, successively called by the DE algorithm to evaluate candidate solutions. The use of the ANN as an approximation model to the physical system allows for the fast and easy testing of different optimization scenarios, concerning different optimization constraints, as the ANN does not need any retraining. Consequently, different operational strategies can be evaluated at a minimal computational cost and the trade-offs between maximization of pumping rates and the minimization of environmental effects can be considered in a more rational and systematic way.

## 2 Area of Study

The area of interest is placed in the Northern part of Rhodes Island in Greece and covers approximately 217km$^2$ (Fig. 1). The extremely high water demand, especially during the summer period, is mainly covered by pumping of subsurface water resources at inland locations, where the water quality is still at high level. The geological formations at the area of interest are primarily limestone, clastic formations, alluvial deposits and sea sands along the coastline [10]. The aquifer depth at the shoreline is 140m and goes up to 400m towards inland.

A 3-d finite-element simulation model of the area of interest has been developed, using the Princeton Transport Code (PTC), a groundwater flow simulator that solves numerically a system of partial differential equations to accurate represent the groundwater flow and the velocities and the contaminant mass transport of the simulated physical system [11]. The aquifer was divided in three layers in order to model the pumping activity of the wells from different depths. The calibration of the model was based on real data (hydraulic heads) from 23 well locations observed in the period 1997-98. The parameters used to calibrate the PTC model were hydraulic conductivity and porosity; the medium was considered homogeneous over the area of interest. The calibration and verification of the simulation model is described in detail in [12].

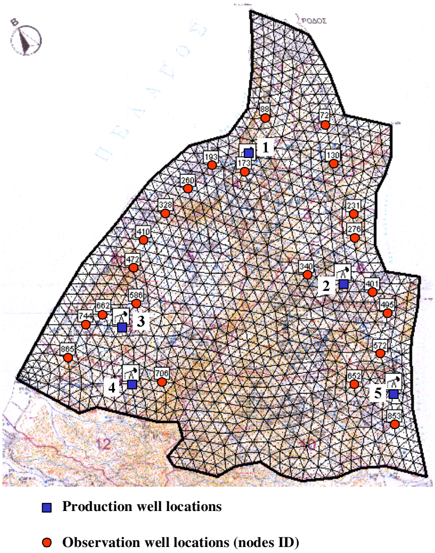


**Fig. 1.** Area of study and its discretization

## 3 The Adopted Methodology

Only five representative pumping wells, with the highest pumping rates, were considered in this study (Fig. 1). The water management problem is considered as an optimization one; its objective is the maximization of the total pumping rate from the five pre-selected pumping locations, without violating the hydraulic head constraints imposed at specific observation locations along the coastal region of the Island (Fig. 1). The actual maximum pumping rates (during the summer period) for the five pumping locations vary between 2200 $m^3$/d and 3880 $m^3$/d. A maximum permitted value of 4000 $m^3$/d for all 5 pumping wells was adopted, to take also into account the (much smaller) pumping rates of neighbouring wells, which are not considered in the present optimization procedure. A minimum value equal to 0 was set for all five pumping wells.

Implicit constraints are related to the minimum permitted hydraulic head at the observation locations near the coastal region; they were taken into account through special penalty functions, included in the cost function of the optimization procedure. The proposed methodology provides the ability to run different scenarios, concerning the imposed constraints, with a minimum computational effort, as the ANN, used as the evaluation function, does not need any retraining. In order to demonstrate this capability, different sets of minimum allowed hydraulic heads were successively used in the optimization procedure. These sets of constraints were calculated in the interval between a) the values of the hydraulic heads at the corresponding observation locations for zero pumping rates and b) those values for the corresponding maximum pumping rates (4000 $m^3/d$). The limiting values for the hydraulic heads (for zero and for maximum pumping rates) were obtained using the PTC numerical model. The cost function $f$ to be minimized is formulated as the sum of three terms:

$$f = f_1 + f_2 + f_3\,,\ f_1 = \frac{20000 - \sum_{i=1}^{5} Q_i}{10000}\,,\ f_2 = \sum_{j=1}^{22} g_j\,,\ f_3 = \begin{cases} 1 & if\ f_2 > 0 \\ 0 & otherwise \end{cases} \tag{1}$$

$$g_j = \begin{cases} \mathrm{min\_}water\_elev_j - water\_elev_j & if\ \ water\_elev_j < \mathrm{min\_}water\_elev_j \\ 0 & otherwise \end{cases} \tag{2}$$

The first term ($f_1$) is related to the objective of the optimization problem and is designed to maximize the sum of the pumping rates. In the case where all 5 pumping rates take their maximum allowed values (4000 $m^3/d$) term $f_1$ takes its minimum value (equal to zero). The second term ($f_2$) is a penalty term, designed to materialize the implicit constraints of the optimization problem. This term takes a zero value if no constraint is violated otherwise it varies linearly with the magnitude of constraint violation. The third term ($f_3$) is an additional penalty term, which ensures that all non-feasible solutions (which violate the constraints) have a higher cost function than the feasible ones. Terms $f_2$ and $f_3$ become zero for all feasible solutions.

### 3.1 Outline of the Procedure

In order to evaluate each candidate solution (a set of pumping rates), during the optimization procedure, the calculation of the hydraulic heads in each one of the observation locations is needed, so that the constraints can be evaluated and the cost function is then computed. However, each call to the numerical simulator (PTC algorithm) is a time consuming procedure (than needs about 2.4 minutes in a Pentium M, 1.73 GHz Notebook), and for successive calls to the numerical simulator (during the optimization procedure) the computation time becomes impractical. The computation time explodes if many different scenarios are necessary for optimization purposes (i.e. different minimum allowed the hydraulic heads at the observation locations).

In this work the time consuming calls of the PTC algorithm are replaced with an ANN, properly trained to adequately simulate the physical system under consideration. As the computation time of a call to a trained ANN is negligible, compared to a call to the PTC simulator, an optimization run performed with an Evolutionary

Algorithm (EA) linked to the ANN can be performed in less than a 5 minutes. On the contrary, an optimization run using an EA and the PTC numerical simulator needs tenths of hours to be completed (about 160 hours in a Pentium M, 1.73 GHz Notebook). In our case the ANN is used as an approximation model to the "precise" PTC numerical simulator; consequently, the resulting optimal solution becomes accurate and meaningful as far as a very good approximation to the physical system is provided by the ANN.

The adopted ANN is a classic fully connected multi-layer perceptron (MLP), with a single hidden layer, trained in a supervised manner with the error back-propagation algorithm. A more sophisticated training algorithm would reduce the training time, but this was a very small percentage of the whole training procedure, the large part being the PTC computations for the training set. The number of nodes in the input layer is equal to the number of pumping wells (5), while the number of nodes in the output layer is equal to the number of observation wells (22). The activation function is the commonly used logistic function. More details about the set up and the training of the ANN, and the adopted procedure can be found in [12].

During the optimization procedure the computation time for each evaluation of the ANN is negligible; however, the training time of the ANN is considerably high. The ANN needs a lot of training data, which in our case are sets of input pumping rates at the corresponding pumping wells and sets of the hydraulic heads at the corresponding observation locations. The training sets were obtained using the PTC numerical simulator. The effective range of each pumping rate (0 - 4000 $m^3/d$) was divided into four intervals of 1000 $m^3/d$, resulting in five different pumping rate values (0, 1000, 2000, 3000, and 4000 $m^3/d$) which yields $5^5 = 3125$ possible pumping sets for all five pumping wells (computed in approximately 130 hours). The PTC simulator was used to provide 3125 sets of input pumping rates and output the hydraulic heads for training the ANN. After the training, the ANN is used in a way similar to a mathematical function to provide the hydraulic heads at the 22 observation locations for every set of pumping rates in the range between 0 and 4000 $m^3/d$.

A Differential Evolution algorithm [13] was adopted as the optimization methodology to provide the optimal set of pumping rates that, additionally, fulfill the imposed constraints. The classic DE algorithm evolves a fixed size population, which is randomly initialized. After initializing the population, an iterative process is started and, at each generation, a new population is produced until a stopping condition is satisfied. At each generation, each element of the population can be replaced with a new generated one. The new element is a linear combination between a randomly selected element and a difference between two other randomly selected elements.

The DE optimizer used in this work enables the external use of executable or batch files for evaluation purposes and can be applied without modifications to various design optimization problems [14]. Two executables are used to evaluate the cost function for each candidate solution. First, the ANN (working in evaluation mode) is called to compute the hydraulic heads. Then, a second executable is called to compute the cost function (Eq. 1-2) for the corresponding candidate solution. In order to run a different scenario, for a different set of constraints, a different file for the permissible hydraulic heads can be used, and thus a different cost function is evaluated, without the need for retraining the ANN.

## 4 Results and Discussion

Eleven different optimization scenarios were considered, to demonstrate the ability of the proposed procedure to provide optimal solutions for different sets of environmental constraints at a minimal computational cost.

**Table 1.** PTC calculations of the hydraulic heads, at the 22 observation wells for the two extreme cases of a) all zero pumping rates and b) equal to the maximum allowed

| Observation location | Grid node | PTC computed values for zero pumping rates (m) | PTC computed values for maximum pumping rates (m) |
|---|---|---|---|
| 1 | 865 | 160.6 | 158.72 |
| 2 | 744 | 162.56 | 159.22 |
| 3 | 622 | 161.71 | 158.89 |
| 4 | 472 | 165.39 | 163.9 |
| 5 | 410 | 158.25 | 157.68 |
| 6 | 328 | 154.96 | 154.59 |
| 7 | 260 | 152.26 | 151.48 |
| 8 | 193 | 149.61 | 147.56 |
| 9 | 88 | 147.32 | 146.03 |
| 10 | 72 | 148.75 | 148.09 |
| 11 | 130 | 155.34 | 154.45 |
| 12 | 231 | 151.94 | 151.64 |
| 13 | 276 | 153.45 | 153.25 |
| 14 | 401 | 157.83 | 157.58 |
| 15 | 495 | 155.44 | 155.05 |
| 16 | 572 | 163.51 | 161.99 |
| 17 | 853 | 155.7 | 153.93 |
| 18 | 652 | 180.57 | 178.44 |
| 19 | 340 | 180.98 | 180.61 |
| 20 | 173 | 157.68 | 149.63 |
| 21 | 586 | 180.69 | 176.32 |
| 22 | 706 | 208.84 | 204.94 |

These scenarios correspond to eleven different sets for the limiting values of the hydraulic heads at the 22 observation locations. These eleven sets were computed by taking a linear interpolation (with a step of 10%) between two extreme sets of hydraulic heads; the first set corresponds to the hydraulic heads at the observation locations for zero pumping rates for all pumping wells; the second set corresponds to the hydraulic heads for the maximum pumping rate (4000 $m^3$/d) for all five pumping wells. These two extreme sets of values were obtained using the PTC numerical simulator, and are listed in Table 1. All runs of the DE optimization algorithm were performed with the following DE parameters: $F$=0.9, $C_r$=0.7, number of generations = 200, population size= 20.

Fig. 2 contains the results of the eleven scenarios (optimization procedures), corresponding to the different sets of hydraulic heads at the observation locations. For all

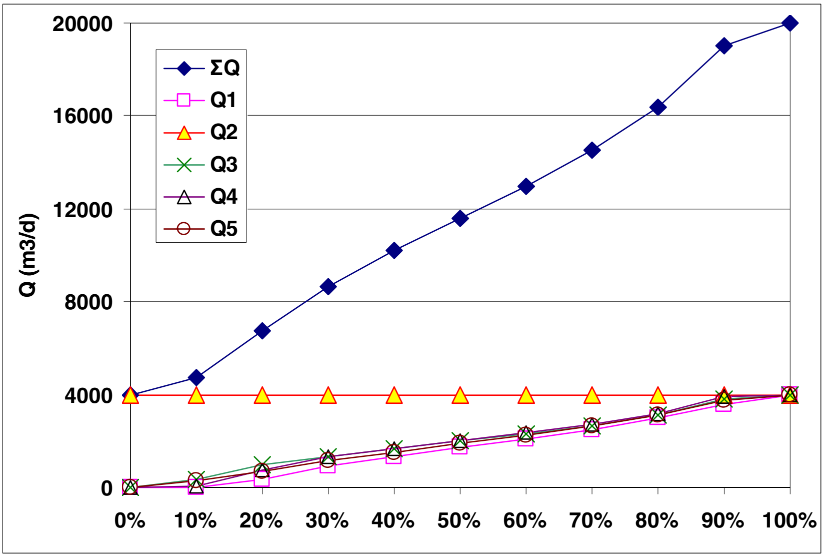


**Fig. 2.** The results of the eleven optimization procedures for the corresponding sets of constraints

but the second production wells (including the total pumping rate) a non linear variation of the pumping rates is observed. The value of the pumping rate for the second production well is equal to the maximum one for all cases considered, which means that its effect on the hydraulic heads on the corresponding locations is minimal. The results contained in Fig. 2 can be used as a guide to decide on the best water management policy for the region under consideration, by accounting for the trade-offs between maximization of pumping rates and the minimization of environmental effects.

## 5 Conclusions

The procedure described in this work combines a MLP ANN with a DE optimizer to successively provide optimal solutions to a water management problem, for different sets of the environmental constraints, at a minimal computational cost. The objective of the optimization procedure is to maximize the total pumping rate of the pumping wells, without violating the environmental constraints associated with the water table drawdown at pre-specified locations. Although the training of the ANN necessitated a large number of costly runs of the PTC groundwater flow simulator, after its training it can be used without modification for several optimization runs, for different sets of constraints, connected to the values of the limiting water table drawdown at the specified locations. As a result, much lower computational time is needed for multiple optimization runs, compared to the case where the optimization algorithm was directly combined with the groundwater flow simulator, based on a finite element model, and useful information can be provided to assist the decision for the best management policy.

## References

1. Freeze, R.A., Cherry, J.A.: Groundwater. Prentice-Hall, Inc., New Jersey (1979)
2. Rogers, L.L., Dowla, F.U.: Optimization of Groundwater Remediation Using Artificial Neural Networks With Parallel Solute Transport Modeling. Water Resources Research 30, 457–481 (1994)
3. Rogers, L.L., Dowla, F.U., Johnson, V.M.: Optimal Field-Scale Groundwater Remediation Using Neural Networks and the Genetic Algorithm. Environmental Science and Technology 29, 1145–1155 (1995)
4. Morshed, J., Kaluarachchi, J.J.: Application of Artificial Neural Network and Genetic Algorithm in Flow and Transport Simulations. Advances in Water Resources 22, 145–158 (1998)
5. Wen, C.-G., Lee, C.-S.: A Neural Network Approach to Multiobjective Optimization for Water Quality Management in a River Basin. Water Resources Research 34, 427–436 (1998)
6. Aly, A.H., Peralta, R.C.: Comparison of a Genetic Algorithm and Mathematical Programming to the Design of Groundwater Cleanup Systems. Water Resources Research 35, 2415–2425 (1999)
7. Rao, S.V.N., Thandaveswara, B.S., Murty Bhallamudi, S., Srinivasulu, V.: Optimal Groundwater Management in Deltaic Regions Using Simulated Annealing and Neural Networks. Water Resources Management 17, 409–428 (2003)
8. Arndt, O., Barth, T., Freisleben, B., Grauer, M.: Approximating a Finite Element Model by Neural Network Prediction for facility Optimization in Groundwater Engineering. Europ. J. of Oper. Research 166, 769–781 (2005)
9. Yan, S., Minsker, B.: Optimal Groundwater Remediation Design Using an Adaptive Neural Network Genetic Algorithm. Water Resources Research 42 (2006)
10. Papadopoulou, M.P., Vondikaki, E., Stergiadi, M., Karatzas, G.P.: Optimal Fresh Water Management with Emphasis in Environmental Quality Constraints in the Northern part of Rhodes Island, Greece. In: VIII Protection and Restoration of the Environment Conference, Chania, Greece (2006)
11. Babu, D.K., Pinder, G.F., Niemi, A., Ahlfeld, D.P., Stothoff, S.A.: Chemical Transport by Three-Dimensional Groundwater Flows, 84-WR-3. Princeton University, USA (1997)
12. Nikolos, I.K., Stergiadi, M., Papadopoulou, M.P., Karatzas, G.P.: Artificial Neural Networks as an Alternative Approach to Groundwater Numerical Modelling and Environmental Design. Hydrological Processes (2008)
13. Price, K.V., Storn, R.M., Lampinen, J.A.: Differential Evolution, a Practical Approach to Global Optimization. Springer, Heidelberg (2005)
14. Nikolos, I.K.: Inverse Design of Aerodynamic Shapes using Differential Evolution coupled with Artificial Neural Network. In: Proceedings of the ERCOFTAC Conference in Design Optimization: Methods and Applications, Athens (2004)

# Using Hierarchical Task Network Planning Techniques to Create Custom Web Search Services over Multiple Biomedical Databases

Miguel García-Remesal

Biomedical Informatics Group, Dep. Inteligencia Artificial, Facultad de Informática,
Universidad Politécnica de Madrid. Campus de Montegancedo S/N, 28660 Boadilla del Monte,
Madrid, Spain
mgarcia@infomed.dia.fi.upm.es

**Abstract.** We present a novel method to create complex search services over public online biomedical databases using hierarchical task network planning techniques. In the proposed approach, user queries are regarded as planning tasks (goals), while basic query services provided by the databases correspond to planning operators (POs). Each individual source is then mapped to a set of POs that can be used to process primitive (simple) queries. Advanced search services can be created by defining decomposition methods (DMs). The latter can be regarded as "recipes" that describe how to decompose non-primitive (complex) queries into sets of simpler subqueries following a divide-and-conquer strategy. Query processing proceeds by recursively decomposing non-primitive queries into smaller queries, until primitive queries are reached that can be processed using planning operators. Custom web search services can be created from the generated planners to provide biomedical researchers with valuable tools to process frequent complex queries.

**Keywords:** Database integration, automated planning, hierarchical task network planning.

## 1 Introduction

Over the last few years, there has been a dramatic increment in the number of publicly available biomedical databases [1]. The latter provide information on complementary topics, including biomedical literature, diseases, genes, proteins, polymorphisms, etc.

Public online resources are normally focused on single topics. For instance, PubMed [2] provides references to literature, while OMIM [3] is a database of genetic disorders. Therefore, to process frequent queries that involve searching for several topics, users are required to manually follow ad-hoc search flows that define chained sequences of searches on different databases. An example of such queries would be "retrieve all European laboratories that perform diagnostic tests for the Cystic Fibrosis disease". This search would entail the use of the following search flow: i) searching the OMIM database to get the disease identifier (MIM ID) associated to that particular

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 42–49, 2008.

disease, and ii) querying the EDDNAL [4] database using the previously obtained MIM ID to retrieve all the laboratories that perform clinical tests for the target genetic disorder. Manually executing these search flows becomes a harder task when the number of involved databases increases. Therefore, there is a need for novel methods and tools to automatically perform these complex searches.

In this paper we propose the use of hierarchical task planning (HTN) techniques [5, 6] to facilitate the creation of complex query services over public online biomedical databases. In our approach, user queries are regarded as tasks to be performed—i.e. goals to be achieved—while basic query services provided by the sources are viewed as primitive planning operators (POs). For instance, the PubMed biomedical literature database provides different basic query services—e.g. searches by topics, authors, journals, etc. Thus, a user query such as “find all articles by John Doe” could be considered as a task (or goal) to be achieved. Similarly, the query service “search all the articles by a concrete author” provided by PubMed could be regarded as a PO that can be used to carry out that particular task.

The idea behind the proposed method is exploiting the domain knowledge provided by the search flows followed by users to create query decomposition methods (DMs) to deal with non-primitive queries. The latter are complex queries that involve a chained sequence of searches in one or more databases. Unlike simple queries, non primitive queries cannot be processed by applying a single primitive planning operator. Instead, they require the execution of a sorted sequence of primitive operators. Hence, DMs can be defined as pieces of domain knowledge that describe how to recursively decompose complex queries into a set of simpler queries following a divide-and-conquer strategy. These subqueries can be either primitive or non-primitive. Further decomposition is then performed on non-primitive subqueries until primitive queries are reached that can be processed using individual planning operators.

Once a planner based on specific DMs extracted from a given search flow has been created, it can be used as the core of a web search service that supports that concrete search flow. The generated web services can be either used as independent tools, or can be reused as building blocks to create more complex query services.

This paper is organized as follows. In the next section, we focus on the methods we used to create web search services from search flows using HTN planning techniques. Next, we present the results of an experiment that we conducted to test our approach. The experiment involved an actual search flow frequently used in the domain of genetic diseases. After that, we briefly compare our method with other similar approaches. Finally, we draw the conclusions.

## 2 Methods

The first step toward the creation of custom web services supporting concrete search flows is the identification of the involved databases. Once the sources have been identified, each database is then mapped to a set of planning operators (POs) describing the basic search services provided by that particular source. To represent these operators, we use the classical representation [7] for planning problems, based on first order logic (FOL).

A PO can be defined as a tuple *o* = (name(*o*), pre(*o*), del(*o*), add(*o*)) whose elements are as follows.

- name(*o*) represents the name of the operator. It is a syntactic expression of the form n($x_1$: $t_1$, $x_2$: $t_2$, …, $x_k$: $t_k$), where n is a unique operator symbol, and $x_1$, …, $x_k$ are all variable symbols that represent the operator's parameters. The symbols $t_1$, …, $t_k$ represent the types associated to the different variables. Operators can be instantiated by binding one or more variables in name(*o*) to constant values belonging to their corresponding types.
- pre(*o*) is a set of literals—i.e. FOL atoms—that represent the precondition of the operator *o*. An instance of *o* is said to be applicable iff its precondition holds in the current state.
- del(*o*) is the set of negative effects of the operator, also called the *deletion list*. It is a set that includes all the atoms that will no longer hold after the execution of the operator.
- add(*o*) is the set of positive effects of the operator, also called the *addition list*. It contains all the atoms that will hold after the execution of the operator.

To illustrate the translation of basic query services into POs, let us consider the PubMed database. As stated previously, PubMed provides the service "search all the articles by a given author". The PO associated to this query service can be defined as follows.

**PUBMED_QUERY_OP_search_by_author(?q: string)**

**pre:** AUTHOR_QUERY_STRING(?q)
**del:** AUTHOR_QUERY_STRING(?q)
**add:**{article_by(?p ?a) | paper ?p is authored by author ?a}

As shown above, the operator *PUBMED_QUERY_OP_search_by_author* takes as input the variable *?q* of type *string*. Using this parameter we indicate the author of the papers that we are interested in retrieving. Regarding the operator's precondition, it states that the atom *AUTHOR_QUERY_STRING(?q)* must hold in the current state for the operator to be applicable. This atom is automatically asserted when the user launches the query. After the search has been completed, the atom *AUTHOR_QUERY_STRING(?q)* is no longer needed, and thus, it is discarded. Besides, a set of instances of the atom *article_by(?p ?a)* are automatically asserted, thus holding in the next state. This set includes all instances of the atom *article_by(?p ?a)* such that the paper *?p* has been published by the author *?a* according to the records retrieved from PubMed. Note that more than one autor *?a* might match the user query *?q*.

Apart from creating operators specifically designed to process user queries, it is also necessary defining POs that enable the planning algorithm to achieve intermediate goals. An example is provided next.

**PUBMED_OP_search_by_author(?pn: person)**

**pre:** person(?pn)
**del:** person(?pn)
**add:** {article_by(?p ?pn) | paper ?p is authored by author ?pn}

The PO shown in the above example is similar to the operator *PUBMED_QUERY_OP_search_by_author*. Indeed, both POs provide the same functionality. The only difference is that the operator's precondition must be asserted by a previously applied operator rather than by a user query.

Once all the target databases have been mapped to sets of primitive POs, it is now possible to encode a query processing task as a planning problem.

The corresponding planning problem is defined by the tuple $P(O, s_0, g)$, whose elements are as follows.

- $O$ is a set of operators. This includes all POs provided by the involved sources.
- $s_0$ is the initial state of the world. It is represented by a set that includes all the atoms corresponding to the query launched by the user. For instance, the user query *"retrieve all articles authored by J. Doe"* would be represented by the initial state $s_0$ = *{AUTHOR_QUERY_STRING("J. Doe")}*.
- $g$ is the goal state. It is represented by a set that includes all atoms that must hold in the goal state. Going back to the previous example, the goal represented by the set $g$ = *{article_by(?a "J. Doe")}* includes all instances of the predicate *article_by* such that "J. Doe" is the author of the article *?a*.

Once the planning problem has been defined, it is now possible to extract a solution plan $\Pi = \{\pi_1, \pi_2, \ldots, \pi_m\}$ using any automated planning method [7]. Note that each $\pi_i \in \Pi$ represents a search in one concrete database that takes as input some of the results provided by previously executed searches—i.e. $\pi_1, \ldots, \pi_{i-1}$.Thus, $\Pi$ represents a query execution plan that can be automatically executed.

The main drawback of this approach is that classical planning techniques do not exploit the expert knowledge provided by the query flows followed by users to manually execute complex frequent queries.

Among all available automated planning techniques, we believe that hierarchical task network planning techniques (HTN) [5, 6] are the best suited to address the creation of custom web search services. This is partly because HTN provides specific artifacts called decomposition methods (DMs) that can be regarded as "recipes" that describe how a human expert may think about manually processing a complex query.

Hence, we propose translating the query flows used by human experts to process frequent queries into DMs built upon primitive operators provided by the databases. These DMs i) facilitate the planning task, and ii) generate web search services that are similar to how human experts manually execute the queries.

DMs can be regarded as methods to solve complex tasks that can be recursively divided into sets of simpler tasks following a divide-and-conquer strategy. For instance, the query "find all European laboratories that perform diagnostic tests for the Cystic Fibrosis disease" is handled by human experts by decomposing it into two simpler queries. First, the user queries the OMIM database to obtain the disease identifier (MIM ID) associated to the target disease (i.e. Cystic Fibrosis). Next, the EDDNAL database is searched by providing the previously retrieved MIM ID as input. This produces a result set including all European laboratories that perform genetic tests for the Cystic Fibrosis genetic disorder. This search flow can be easily translated into a set of DMs and primitive operators as shown below.

**METHOD_search_for_lab_by_disease(?q: string)**
  **task:** search_for_lab_by_disease(?q: string)
  **pre:** none
  **del:** none
  **subtasks:** <OMIM_QUERY_OP_get_matched_diseases(?q),
    get_laboratories(?m)>

**METHOD_get_laboratories()**
  **task:** get_laboratories()
  **pre:** none
  **del:** none
  **subtasks:** <EDDNAL_OP_get_laboratory(?m), get_laboratories()>

**OMIM_QUERY_OP_get_matched_diseases(?q: string)**
  **pre:** DISEASE_QUERY_STRING(?q)
  **del:** DISEASE_QUERY_STRING(?q)
  **add:** {disease-id(?m ?d) | ?m is the identifier associated to disease[1] ?d according to OMIM}

**EDDNAL_OP_get_laboratories(?m: disease-id)**
  **pre:** disease-id(?m ?d)
  **del:** disease-id(?m ?d)
  **add:** {test-lab(?l ?m ?d) | ?l is a lab that performs tests for disease ?d according to EDDNAL}

When using HTN techniques, goals are no longer represented as sets of atoms that must hold in the goal state. Instead, goals are regarded as lists of tasks to be performed. Thus, the statement of a HTN planning problem is represented by the tuple $P = (O, M, s_0, t)$, where $M$ is a set that includes all the DMs, $t$ is a list of tasks to be achieved, and $O$ and $s_0$ have the same meaning as in classical planning.

Using the above definitions, the query flow "retrieve all European laboratories that perform diagnostic tests for the Cystic Fibrosis disease" can be stated as the following planning problem *P = {O, M, s₀ = {DISEASE_QUERY_STRING("Cystic Fibrosis")}, t = <search_for_lab_by_disease(?q)>}*.

This HTN planning problem can be solved using the SHOP2 HTN planning algorithm [8]. SHOP2 proceeds by recursively searching for a suitable method or operator to achieve the goal task. In the previous example, the task *search_for_lab_by_disease* can be performed by applying the method *METHOD_search_for_lab_by_disease*. Once the method has been applied, the task *search_for_lab_by_disease* is replaced with the corresponding list of subtasks specified by selected method, i.e. *<OMIM_QUERY_OP_get_matched_diseases, get_laboratories(?m)>*. Note that the first subtask can be directly achieved by a primitive operator, while, the second is a non-primitive subtask that requires further decomposition using a suitable method—i.e. the method *METHOD_get_laboratories*.

[1] Note that more than one disease *?d* might match the query string *?q*.

Once the search flow has been converted into a HTN planning problem, we then use a modified version of the JSHOP2 planner generator [9] to create a web search service. The modified planner generator takes as input the formal definition of the HTN problem augmented with additional information. The latter includes directions on how to enter and extract information from the web pages belonging to the different databases. The planner generated by JSHOP2 is then used as the core of the newly created web service, acting as a mediator responsible for query processing.

## 3 Results

In this section, we present the results of an experiment that we carried out to test the proposed approach. The experiment involved a complex search flow frequently used by biomedical researchers on the area of genetic diseases. The search flow is depicted in the figure below.

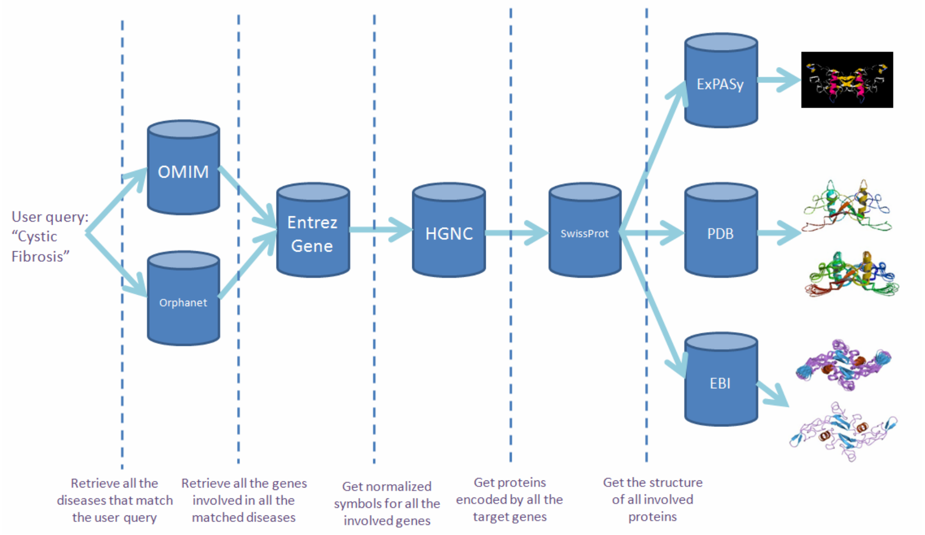

User query: "Cystic Fibrosis"
OMIM
Orphanet
Entrez Gene
HGNC
SwissProt
ExPASy
PDB
EBI
Retrieve all the diseases that match the user query
Retrieve all the genes involved in all the matched diseases
Get normalized symbols for all the involved genes
Get proteins encoded by all the target genes
Get the structure of all involved proteins

**Fig. 1.** Overview of the search flow involved in the experiment

As shown in figure 1, the search flow is aimed to retrieve the three-dimensional structure of all proteins involved in a given genetic disease. This generic query entails a complex chained search over 8 databases focused on different topics. These topics include diseases (OMIM [3]), rare genetic diseases (Orphanet [16]), genes (Entrez Gene [17]), normalized gene symbols (HGNC [18]), proteins (SwissProt [19]), and protein structures (ExPASy [19], PDB [20] and EBI [21]). The different phases of the search are shown in the figure.

We encoded the search flow as an equivalent HTN planning problem using the methods described in the previous section. The obtained HTN problem included 8 POs and 5 DMs.

Once the search flow was encoded as a HTN planning problem, we used the JSHOP2 planner generator to automatically create a HTN planner implementing the target search flow. The generated planner was then encapsulated by a Java web service that was deployed in an application server.

We evaluated the performance of the generated web service by launching a set of queries related to 200 genetic diseases. Service times ranged between 3 and 12 minutes using a server based on Windows Vista Ultimate™ with 4 GB of RAM. These timings include both the plan generation and the retrieval of all structures belonging to all proteins involved in the target genetic disorder.

We believe that the generated web service facilitates the execution of the corresponding search flow, considering that manually retrieving the 3D structure associated to a single protein takes in average 2 minutes.

In the next section we compare the proposed method to other existing approaches to integrate public online biomedical databases.

## 4 Discussion

In recent years, different approaches have been proposed in the literature to address the integration of web-based biomedical databases. This includes information linkage [10], mediator/wrapper based methods [11], ontology-based mediation approaches [12, 13], and automated planning techniques [14]. We believe that planning techniques are particularly well suited to integrate public online databases since, i) they can process queries not supported by information linkage-based methods, and ii) unlike mediation-based approaches, they do not require establishing complicated mapping relationships between the schemas of the databases.

The BACIIS system [14], based on classical planning methods, relies on a custom ontology called BaO that includes all the relevant classes of objects in the domain together with a set of relationships that represent the inputs/outputs accepted/provided by the different databases. Queries are executed by generating a query processing plan—i.e. a sequence of searches on different databases—using the information provided by the ontology. Plans are automatically created using a modified version of the domain-independent GraphPlan [15] planning algorithm. The main drawback of this approach is that domain knowledge—i.e. the search flows—used by experts to manually process complex queries is not exploited to facilitate the creation of the plans.

Conversely, our method exploits the expert knowledge provided by the search flows to generate web search services that are similar to how human experts manually execute the queries. Besides, the DMs implemented by previously created web services can be reused as components supporting complex query services.

## 5 Conclusions

In this paper, we propose the use of HTN planning techniques to create complex web search services over multiple public online biomedical databases. HTN planning provides a framework to encode manual search flows used by biomedical researchers as HTN planning problems. Planners created to solve these HTN problems can be used as mediators that perform searches in a similar way as human experts execute queries.

Besides, the generated web services can be reused as components to build more complex query services, thus facilitating the integrated access to multiple public online biomedical resources.

## References

1. Galperin, M.Y.: The Molecular Biology Database Collection: 2008 Update. Nucleic Acids Research 36(Database issue), D2–D4 (2008)
2. `http://www.ncbi.nlm.nih.gov/pubmed/` (last accessed, April 2008)
3. `http://www.ncbi.nlm.nih.gov/sites/entrez?db=omim` (last accessed, April 2008)
4. `http://www.eddnal.com/` (last accessed, April 2008)
5. Sacerdoti, E.: The nonlinear nature of plans. In: Proceedings of the International Joint Conference on Artificial Intelligence, pp. 206–214 (1975)
6. Erol, K., Hendler, J., Nau, D.: Semantics for hierarchical task-network planning. Technical Report CS TR-3239, UMIACS TR-94-31, ISR-TR-95-9. University of Maryland (1994)
7. Ghallab, M., Nau, D., Traverso, P.: Automated Planning: Theory and Practice. Morgan Kaufmann, San Francisco (2004)
8. Nau, D., Au, T.C., Ilghami, O., Kuter, U., Murdock, J.W., Wu, D., Yaman, F.: SHOP2: An HTN Planning System. Journal of Artificial Intelligence Research 20, 379–404 (2003)
9. `http://www.cs.umd.edu/projects/shop/description.html` (last accessed, April 2008)
10. Dias, G., Oliveira, J.L., Vicente, F., Martín-Sánchez, F.: Integrating Medical and Genomic Data: a Successful Example of Rare Diseases. Stud. Health Technol. Inform. 124, 125–130 (2006)
11. Haas, L.M., Schwarz, M., Kodali, P., Kotlar, E., Rice, J.E., Swopre, W.C.: DiscoveryLink: A system for Integrated Access to Life Sciences Data Sources. IBM Systems Journal 40(2), 489–511 (2001)
12. Stevens, R., Baker, P., Bechhofer, S., Ng, G., Jacoby, A., Paton, N.W., Goble, C.A., Brass, A.: TAMBIS: Transparent Access to Multiple Bioinformatics Information Sources. Bioinformatics 16(2), 184–186 (2000)
13. Alonso-Calvo, R., Maojo, V., Billhardt, H., Martín-Sánchez, F., García-Remesal, M., Pérez-Rey, D.: An agent- and ontology-based system for integrating public gene, protein and disease databases. Journal of Biomedical Informatics 40(1), 17–29 (2007)
14. Miled, Z.B., Li, N., Bukhres, O.: BACIIS: Biological and Chemical Information Integration System. Journal of Database Management 16(3), 72–85 (2005)
15. Blum, A., Furst, M.: Fast Planning Through Planning Graph Analysis. Artificial Intelligence 90, 281–300 (1997)
16. `http://www.orpha.net/consor/cgi-bin/index.php` (last accessed, April 2008)
17. `http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene` (last accessed, April 2008)
18. `http://www.genenames.org/` (last accessed, April 2008)
19. `http://www.expasy.ch/sprot/` (last accessed, April 2008)
20. `http://www.rcsb.org/pdb/home/home.do` (last accessed, April 2008)
21. `http://www.ebi.ac.uk/msd/` (last accessed, April 2008)

# Building an Index of Nanomedical Resources: An Automatic Approach Based on Text Mining

Stefano Chiesa, Miguel García-Remesal, Guillermo de la Calle,
Diana de la Iglesia, Vaida Bankauskaite, and Víctor Maojo

Biomedical Informatics Group, Dep. Inteligencia Artificial, Facultad de Informática,
Universidad Politécnica de Madrid, Spain

**Abstract.** Nanomedicine is an emerging discipline aimed to applying recent developments in nanotechnology to the medical domain. In recent years, there has been an exponential growth of the number of available nanomedical resources. The latter are aimed to different tasks and include databases, nanosensors, implantable materials, etc. This leads to the necessity of creating new methods to automatically organize such resources depending on their provided functionalities. In this paper we will first present a brief overview on the nanomedical discipline and its related technologies. Next we will introduce a method targeted to the automated creation of an index of nanomedical resources. This method is based on an existing approach to automatically build an index of biomedical resources from research papers using text mining techniques. We believe that such an index would be a valuable tool to foster the research on nanomedicine. This is an example of application in the new area of Nanoinformatics.

**Keywords:** nanomedicine, text mining, biomedicine, nanoinformatics.

## 1 Introduction

According to the National Nanotechnology Initiative (NNI), "nanotechnology is the understanding and control of matter at dimensions between approximately 1 and 100 nanometers, where unique phenomena enable novel applications" [1]. In 1959, Richard Feynman presented a conference called "There's plenty of room at the bottom"[2]. In his talk, he touched topics that in a few decades became one of the main objectives for the research world. He did not only predict the progressive miniaturizing process that led to the possibility of work at atomic level, he also considered it crucial that the manufacturing of tools can interact with biologic cells at the same size level.

Nanotechnology actually constitutes a path that leads towards the integration of natural and artificial world. One of its strongest points is the fact that nano-particles and nano-devices can effectively interact with natural organism. Considering this context, from the perspective of computer science, nanotechnology catalyzes the passage from its old domain (the study of phenomena surrounding computers) into a more modern one, in which computer science can be defined as the study of natural and artificial information processes [3].

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 50–57, 2008.

The possibility to interact with the natural biological environment makes the research on the biological processes and the understanding of the information processes that are embedded to them easier. Moreover the possibility to manipulate these information flows opens new research opportunities for artificial information processes.

### 1.1 Nanomedicine Definition

The application of nanotechnology to health care is called nanomedicine. At the end of 2002, The National Institutes of Health (NIH) created a new plan for the study of the nanoscience and nanotechnology applied to the medicine. The European Commission has also shown the same increasing interest in using bio-molecular approaches for the diagnosis, monitoring and treatment of high risk diseases like cancer or cardiovascular conditions and developing micro-nano devices and tools for research and development.

According to Jain [4], nanomedicine is based on three progressively more powerful molecular technologies:

i) Nanoscale-structured materials and devices
ii) Genomics, proteomics and artificially engineered microorganism
iii) Molecular machine systems

These may be used for a large number of application fields that can be organized in the taxonomy [5] can be seen in figure 1.

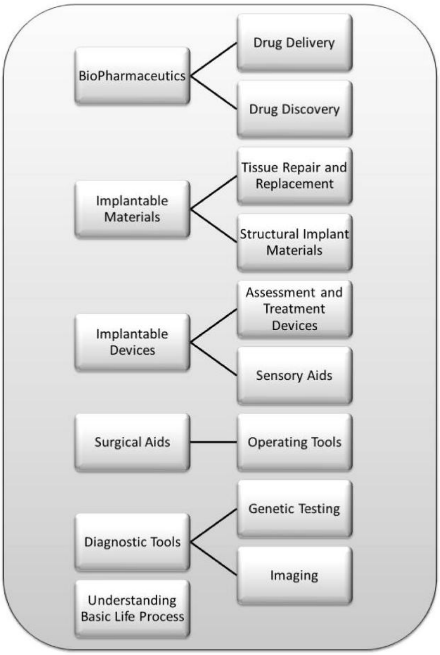


**Fig. 1.** Nanomedicine taxonomy [5]

### 1.2 Nanomedicine Taxonomy

Following Figure 1, nanomedicine can be categorized in several different areas:

**Biopharmaceutics.** This area studies the role of nanotechnology in the pharmaceutical domain. It includes the study of new drugs based on nano-particles (drug discovery) and the nano-systems that can be utilized to deliver the pharmaceutical products in a more effective and precise way (drug delivery).

**Implementable materials.** This area includes the biocompatible materials that can be permanently or temporally implanted in a living organism. These materials can be used for substituting or repairing tissues (tissue repair and replacement) and structures (structural implant materials) of an organism body.

**Implementable devices.** This area contains the technologies that aim towards the creation of nano-devices that can be implanted in live organisms. This category comprises those devices that can process local extracted medical information for diagnosing and treating purposes (assessment and treatment devices). Implantable devices also include devices that can enhance sensory skills restoring lost hearing and sigth functions (sensory aids).

**Surgical Aids.** This area includes the devices that can be helpful for surgical operations. In particular nano tools can be used to perform common surgical tasks in a very precise way or monitoring patient condition with a higher accuracy (operating tools).

**Diagnostic tools.** This area includes the nano-systems that can help to identify the occurrence of a disease as soon as possible. There are possibilities to work directly on genes and genetic samples (Genetic testing) and to create graphical representations that shows images of the patient's condition (imaging).

**Understanding basic life process.** Nanotechnology uses devices and tools created at atomic and molecular sizes. For the biological purposes of understanding life process this is a very important opportunity. Through nanotechnology it is possible to deeply understand the processes like protein folding, and to be able to solve problems that are strongly bound within them.

As shown in this taxonomy there are several areas of interest in medicine, which we will analyze below.

## 2 Background and Rationale

### 2.1 Nanomedicine Overview

The role of nanomedicine over the coming decades will become progressively more central for patient care process. In the "strategic research agenda for nanomedicine" [6] we can find some research lines that are considered the most valuable in respect of benefits for the patient and socio-economical impact.

Nanomedicine will be used for different medical purposes, such as preventive medicine, diagnosis, therapy and follow up monitoring. Current research covers on these aspects of the care process, with respect to different diseases. Nowadays the most frequent cause of death in the European Union is cardiovascular diseases, followed by cancer. Other types of disorders, such as musculoskeletal and inflammatory diseases or neurodegenerative diseases, significantly reduce the patient's quality of life.

For these conditions, nanomedicine will offer solutions to reduce the contraindications that current therapies have, whilst being more effective. Current research is rapidly approaching a stage in which nanotechnologies will be utilized in practice. Methods for fields like biological research and biological imaging, applied to medicine, have been developed using nanosize particles and crystals [7].

Nowadays nanomedicine can count on a huge number of new technologies that can be used in the different branches in which nanomedicine is divided [8]. For instance, surfaces perforated with nanopores are used to create containers that can hide the immunologic system biologic cells belonging to other organisms. Anticancer therapies may count on several nanostructures like fullerenes, nanoshells or tectodendrimers that can perform disease recognitions and target delivering tasks.

In a short period (3 to 5 years) nanotechnology is expected to provide biologic robots to medicine, constituted by engineered bacterial organisms able to produce substances useful for the patient metabolism. They may be used, for example, to increase low levels of vitamins or to produce antitoxins when it is needed.

In a longer period (10 to 20 years) nanomedicine will be able to create artificial red and white blood cells or to replace fragments of chromosomes that may cause genetic diseases to the patient.

The research on nanomedicine is growing fast and the available material about this theme is expected to increase dramatically in the next few years.

### 2.2 Information Management

Over the last few years, the number of papers concerning nanomedicine has increased significantly. The interests on this field has led many research laboratories to increase their effort in obtaining relevant results. The material produced is then shared with the research community.

The rapid growth of produced information leads to organizational issues. Research papers are usually not formally structured. For this reason it is complex to create an automatic approach to collect and order research results. Moreover the possible relation that may exist among different researches is unclear.

This situation may lead to an uncontrolled explosion of hard to manage information. In order to avoid this, automatic approaches to organize the produced material and the existing related concepts (contained in articles) are needed.

Moreover articles may need some standard features in order to structure the information they contain. This may be extremely interesting approach to be able to manage efficiently (automatically) the available resources.

## 3 Methods

### 3.1 Automatic Text Analysis

To address the automated creation of a nanomedical index, we propose a method based on automated text analysis using pattern matching techniques. This method has been previously applied to the biomedical domain to create a tool for automated annotation of biomedical resources from literature [9].

First, linguistic patterns that occur in a text are manually identified using a training set composed of several hundreds of research papers. The extracted patterns describe general syntactic structures and morphological features that allow identification the relevant information to be retrieved from the texts. We considered three different sets of patterns that are used to perform different tasks: i) extracting the names of the resources, ii) identifying their functionalities and iii) classifying the resource into a suitable category.

This classification is driven by the type of resource—e.g. database, annotation tool, visualization tool, etc— and its application domain—e.g. genome, DNA, protein, etc. Types and domains are organized into an ad-hoc taxonomy created by the team of experts.

Using the extracted patterns, we created two transition networks (TN) [10]. These are abstract machines that can determine if a given string belongs to a specific formal language defined by a set of regular expressions.

The first TN extracts the name of the resource and a description of its functionalities, while the second performs the classification task.

The analysis of each document involves 4 stages:

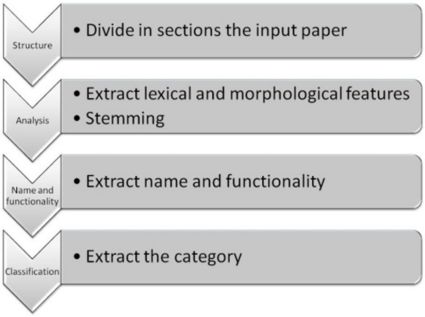

Structure
• Divide in sections the input paper
Analysis
• Extract lexical and morphological features
• Stemming
Name and functionality
• Extract name and functionality
Classification
• Extract the category

**Fig. 2.** Text analysis process

**Structure.** The document is pre-processed to create a surrogate. The latter includes the title, the abstract and a reference to the full text of the article.

**Analysis.** The title and the abstract sections are divided into single sentences that are parsed using a lexical and morphological analyzer. This produces a sequence of tokens labelled with their corresponding linguistic features. Next, the tokens are stemmed, thus being converted into their root form.

**Name and functionality.** The tokens are fed as input of the first transition network to extract the name of the resource and a description of its provided functionalities.

**Classification.** Finally, the tokens are fed as input to the second TN that classifies the tool into the most suitable categories of the taxonomy.

All the extracted information is then stored in a database that can be queried by users through a graphical interface.

### 3.2 Adaptation to the Nano-context

The results obtained in the biomedical domain, encouraged us to explore other possible application fields. We selected the nanomedical domain due to its growing interest in the biomedical community.

First, we identified the required modifications to be performed on the former approach to adapt it to the nanomedicine field.

Similarly as we did for the biomedical domain, we first extracted linguistic patterns by analyzing abstracts of research papers. These linguistic patterns were then used for the creation of the new TNs tuned for the nanomedicine domain. The first issue we had to tackle was that the number of papers on nanomedical resources was much smaller than those related to biomedical resources. This hindered the patterns' extraction process, making it slower and more complex.

The features we were interested in extracting from the documents were the same as for the previous prototype. We wanted to retrieve the resource's name, its functionalities and the category to which it belonged. For some specific resources we also considered which were their required inputs and outputs –e.g. the output of a nanosensor and the input of a diagnostic nanodevice. This can be used to define complex workflows that may involve a sequence of different resources.

### 3.3 Enhance the Text Analysis Performance

The implementation of the proposed text analyzer must address a series of problems. Probably the most difficult is solving the heterogeneity introduced by different researchers when describing their results in research papers. This means that we need a more complex set of linguistic patterns in order to be able to cover all the potential possibilities in which a given concept may be expressed.

This leads us to consider the idea of a standard proposal that may help to enhance the performance of text mining tools. The basic idea consists in encouraging authors to prepare the abstracts of their articles following a predefined structure. The latter defines different patterns that facilitate the extraction of relevant information. This may include, for instance, a set of short sentences introduced by a concrete keyword describing the resource presented in the paper.

## 4 Results and Discussion

The method we described has been tested on the biomedical domain. The set of papers we considered was composed of 392 papers. Among them, a small percentage was related with nanomedicine domain. The first TN managed to recognize, over the whole test set, 376 correct resource names. It also extracts 505 functionalities. The 88% of them were understandable and complete. The 10% were incomplete but still provided useful information. The second TN managed to categorize 305 resources and to assign a domain to 253 resources.

These results refer to the application of the method using TNs mainly tuned for the biomedical field. Our purpose is to refine the set of patterns in order to make them more specific for the nanomedicine domain. The patterns we defined until now are still not stable enough to produce effective results. Nanomedicine is an emerging field and the number of published papers is smaller than in the biomedical domain. High impact journals on nanomedicine are relatively young and is needed a wider set of relevant papers is needed to increase the number of patterns and tune properly our TNs.

## 5 Conclusion and Future Work

We proposed a new approach to the organization of the information produced about nanomedicine. We focussed on a method that led to previous results in the biomedical domain. We analyzed it in order to identify the characteristics of the field of nanomedicine. This opens a path to the implementation of a "nano-resourceome" [11].

We also proposed a way to produce a structured organization of the documents to improve the efficiency of those tools that have to recover information automatically. In this early stage of nanomedicine research, creating the basis for a common standard of articles is a crucial task to be able to track the rapid growth of information about the domain.

The approach we propose is part of the agenda of the ACTION-grid project. The latter is a project supported by the European Commission beginning in June 2008. The main objective is to constitute an international cooperative action on grid computing and biomedical informatics between the European Union, Latin America, the Western Balkans and North Africa. The project includes the survey of the current state of nanotechnology applied to medicine and the production of a White Paper that highlights future research lines based on the synergy between medical informatics and bioinformatics, expanding results towards grid and nano areas (nanotechnology, nanoinformatics, nanomedicine).

In the context of this project the proposed method is a useful tool for information retrieval about nanomedicine resources. These would be categorized and summarized by name and functionalities, giving the researcher a structured index.

Finally we are working towards implementing a software tool able to find relationships between papers in an automatic way. This enhancement provides the possibility of creating workflows among resources or performs more complex queries on the produced nanoresources' database.

**Acknowledgments.** This research has been supported by the European Commission-funded ACGT project (IST-2005-026996) and the Spanish Ministry of Education (TSI2006-13021-C02-01).

## References

1. National Science and Technology Council.: The national Nanotechnology initiative. Strategic Plan. National Science and Technology Council (2007)
2. Faynman, R.P.: There's plenty of room at the bottom. Eng. Sci. 23, 22–36 (1960)

3. Denning, P.J.: Computing is a Natural Science. Commun. ACM 50(7), 13–18 (2007)
4. Jain, K.K.: The handbook of Nanomedicine. Humana Press, Totowa (2008)
5. Gordon, N., Sagman, U.: Nanomedicine Taxonomy. Canadian Institute of Health Research & Canadian NanoBusiness Alliance (2003)
6. European Technology Platform: Nanomedicine: Nanotechnology for Health. European Technology Platform, Brussels (2006)
7. Alivisatos, A.P.: Less is more in medicine. Sci. Am. 285, 66–73 (2001)
8. Freitas, A.R.: Nanomedicine: Nanotechnology, Biology, and Medicine 1(1), 2–9 (2005)
9. García-Remesal, M., Maojo, V., Crespo, J., Billhardt, H.: Logical Schema Acquisition from Text Based Sources for Structured and non-structured Biomedical Sources Integration. In: Proc. AMIA Symp., pp. 259–263 (2007)
10. Woods, W.: Transition Network Grammars for Natural Language Analysis. Commun. ACM 1970 13(10), 591–606 (1970)
11. Cannata, N., Merelli, E., Altman, R.B.: Time to Organize the Bioinformatics Resourceome. PLoS. Comput. Biol. 1(7), 76 (2005)

# Annotation of Colorectal Cancer Data Using the UMLS Metathesaurus

Marcos Martínez, José M. Vázquez, Javier Pereira,
and Alejandro Pazos

IMEDIR Center, University of A Coruña, Spain
{marcosmartinez, jmvazquez, javierp, apazos}@udc.es

**Abstract.** Reusing the data collected during an epidemiological study is a complex task. This is mainly due to the frequent utilization of traditional storage supports, as well as specific formats and/or terminologies which can make it difficult to integrate the gathered data with data from other information resources. Using terms from the UMLS Metathesaurus, we have annotated the data collected during the development of an epidemiological study of colorectal cancer funded by the US National Cancer Institute (NCI), which was carried out in Galicia, Spain. In our opinion, this annotation can facilitate integration of these data with data from other studies, allowing the future development of larger studies in a faster and more economical manner.

**Keywords:** Ontologies, Knowledge Management, Cancer.

## 1 Introduction

Epidemiological studies can be valuable in understanding a number of processes such as the origin and progression of diseases, the detection and the monitoring of outbreaks of pathologies in various populations, and can also aid in decisions regarding the optimization of resources, as well as regarding the overall welfare, safety, and quality of life in general.

However, the study of complex diseases such as different types of cancer, which are caused by a diversity of possibly interrelated genetic, environmental, and lifestyle factors, requires a large amount of data in order to enable researchers to systematically sift through, analyze, and interpret information that is inherently multidimensional, multivalued, heterogeneous, multi-patient, and ideally obtained from the same patients over different periods of time. Moreover, the data must reside in a way that becomes accessible to researchers in a manner that is efficient, reliable, and easy to interpret despite the complex nature of the information itself.

In the case of colorectal cancer, compiling useful and comprehensive data is an expensive and time-consuming task that implies: (1) the capture of a variety of clinical records for every patient included in the study, (2) gathering data and information associated with medical analyses and interpretations for all patients, and (3) conducting personal interviews with each patient and his/her family in order to obtain relevant

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 58–65, 2008.

information, especially information derived from various questionnaires (e.g., risk factor questionnaires).

These valuable data could be reused (totally or partially) by integrating them with data from other similar studies, in such a way that it could be possible to accomplish new studies over a higher volume of information. Nevertheless, these data are normally stored by means of traditional storage supports (e.g. paper) and/or using region-specific terminologies or languages which cannot be understood from other regions in the world. These factors make understanding of underlying semantics a considerably difficult task, what represents an important obstacle in the integration of the gathered data with data from other studies.

An ontological approach can help to facilitate this task. The role of ontologies for allowing a more effective data and knowledge sharing and reusing is widely recognized and, nowadays, there exist a variety of public ontologies from different application domains, which can support the exchange of information among people or systems that use diverse representations to refer to the same meaning.

In an attempt to make a contribution in this area, we have used a well-known set of medical ontologies, the UMLS Metathesaurus [1], to annotate the information collected during an epidemiological study of colorectal cancer in Galicia, Spain. All of this, in order to allow that these data can be understood from other parts of the world and reused in the future to develop larger and more conclusive studies.

## 2 Background

The region of Galicia, located in the northwest of Spain, represents an excellent geographical area for conducting genetic epidemiological studies of colorectal cancer due to several factors: 1) the genetic homogeneity of its population, which facilitates such studies; 2) the relatively high incidence of colorectal cancer; 3) the relatively large size of patients' families, which enhances the statistical significance of studies; 4) the accessibility to the relatives of patients affected by cancer, given that family members often live in the same household or town, and almost always in the same region; 5) the existence of a centralized and universal health care system that allows obtaining a nearly complete case finding and virtually 100% retrieval rate of medical records, pathology reports, tumor blocks, and fresh frozen tissue; 6) a population that appears to be very cooperative with medical studies; 7) the cultural homogeneity of the population, which enables the development of procedures for efficient recruitment of participants in the study and which facilitates the collection of data that are considerably uniform; and 8) the existence of a highly qualified and well motivated group of clinicians and researchers that can fully support the investigations. Combined, these factors make Galicia a rather unique, as well as important, region in which to conduct epidemiological studies in colorectal cancer.

Consequently, these considerations have motivated the development of several research projects in Galicia centered on cancer during the last years, including: (a) "A Pilot Study of colorectal cancer in Galicia, Spain", funded by the US National Cancer Institute (NCI) during the period 2004-2006; (b) "Colorectal Cancer Thematic Network in Galicia", funded by the XUNTA de Galicia during the period 2005-2006; and

(c) “Colorectal Cancer Research Network in Galicia”, funded by the XUNTA de Galicia during the period 2006-2008. Within the framework of these projects, the idea of developing an information system (IS) to improve the traditional way of carrying out epidemiologic studies was jointly raised in 2005 by the Medical Computing and Radiological Diagnosis Center (IMEDIR Center) of the University of A Coruña and The University Hospital Complex Juan Canalejo of A Coruña, which has been funded by the XUNTA de Galicia during the period 2005-2008.

Initially, we developed a secure IS to collect, store and edit medical data through the Internet, from different hospitals [2]. This system was successfully tested by the 4 Galician hospital centers that have participated in the execution of the previously mentioned project: “A Pilot Study of colorectal cancer in Galicia, Spain”. After that, we thought that reusing the data gathered by our IS could provide great benefits, and we started to study the ways of using medical ontologies to semantically tag these data, with the aim of allowing that they can be understood by researchers from all around the world and automatically integrated with other sets of similar data.

The problem of establishing connections between databases and ontologies has already been studied by other authors, who proposed diverse remarkable solutions (e.g. [3] – [7]), based on different approaches [8]. Nevertheless, the case presented in this work has its own peculiarities, which make it different from previous problems: we want to map a fixed set of questions and answers from questionnaires to ontological terms. Some of these questions/answers have a complex semantics and it was necessary to use not only one term from the ontology, but several terms which had to be properly combined in order to express all the meaning contained in each question/answer. The advantage of this situation with respect to other problems is that it only was necessary to find the connections between the elements of each questionnaire and the ontology once, because the same set of questionnaires is used during the entire study.

## 3 Methods

### 3.1 Description of the Collected Data

In the study: “A Pilot Study of colorectal cancer in Galicia, Spain”, a wide range of information about patients affected by colorectal cancer (i.e. probands) and their blood relatives (including parents, siblings and cousins) has been collected. In total, 230 members (62 probands and 168 relatives) from 62 different families have participated in the study. The data collection has been achieved during personal interviews in which the interviewed person has filled out several questionnaires, helped by the medical personnel in hospitals (i.e. doctors or nurses who participated in the study).

The questionnaires used in the study were the following ones: the risk factors questionnaire and the familiar questionnaire, both originally designed by researchers of the NCI and translated to Spanish. The risk factors questionnaire is a comprehensive set of questions about medical history and several aspects of lifestyle (e.g. diet, physical activity, alcohol consumption, etc.), to identify possible risk factors for colorectal cancer. This questionnaire has 421 possible questions (many of which do not have to

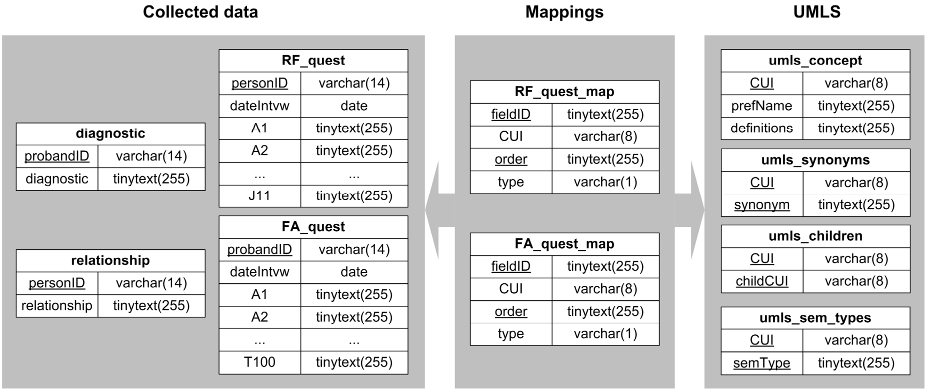


**Fig. 1.** Tables in the database

be answered, because some responses open or close different paths of future questions) and it has been filled out by all the members of the study (i.e. both probands and relatives). On the other hand, the familiar questionnaire has been designed to collect information about the medical history of each family. This questionnaire has 126 possible questions and it only has been filled out by the proband of each family. The filled questionnaires were sent through the Internet from each hospital center to a centralized database (DB), located at the IMEDIR Center of the University of a Coruña, where they were stored.

This centralized DB (see *Collected data* in Fig. 1) contains (1) a table for each kind of questionnaire: the *RF_quest* and *FA_quest* tables, (2) a *diagnostic* table, which stores the diagnostic for each proband and (3) a *relationship* table, which stores the kinship of each member to the proband (e.g. father, sister, proband, etc.). Each row in the *RF_quest* and *FA_quest* tables stores a code that identifies the interviewed person (*personID*), the date of the interview (*dateIntvw*) and the answers of the interviewed person to each question of the questionnaire (column names refer to the code of each question, e.g. A1, B1, B2b, etc.).

For example, let's suppose the following question (extracted from the risk factors questionnaire): *B2. Have you ever had a sigmoidoscopy?* Possible answers are: *(a) Yes*, *(b) No*, *(c) Don't know*. If we suppose that the interviewed choose the answer *(a) Yes*, then the table of the DB that stores the information from the risk factors questionnaire (i.e. *RF_quest*), would store an *a* in the cell located at the column *B2*, at the row corresponding to the interviewed.

However, some answers are more complex, and we had to conceive a representation to store them (see Table 1). An example could be the question *B2a. When was the first time you had a sigmoidoscopy?* Possible answers: *(a) Year*, *(b) Don't know*. Let's

**Table 1.** Format of complex answers

| Format | Description |
|---|---|
| x[Y] | Answer x, with value Y |
| x-y-z | Answer x, y and z (multiple answer) |

suppose that the interviewed answers *a, 1996*. In this case, the row corresponding to this person would store the string *a[1996]* at the column *B2a*.

## 3.2 Mapping the Collected Data to the UMLS Metathesaurus

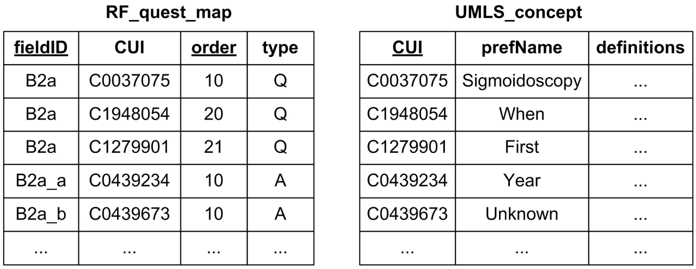
The Unified Medical Language System Metathesaurus (http://umlsks.nlm.nih.gov), developed by the US National Library of Medicine (NLM) is a compilation of names, relationships, and associated information from a variety of biomedical naming systems representing different views of biomedical practice or research. It integrates over 2 million names for some 900 000 concepts from more than 60 families of biomedical vocabularies, as well as 12 million relations among these concepts [1]. One of the vocabularies included in the Metathesaurus is the NCI Thesaurus [9], a broad terminology of the cancer domain built by the NCI to better organize, enhance and leverage the confusing patchwork of clinical and research cancer terms in current use.

Due to these factors, we selected the UMLS Metathesaurus as the most complete and appropriate resource to represent our data. We used version 2007AC of UMLS. With the purpose of navigating rapidly through the parent-child (IS_A) relations (e.g. *Polypectomy* IS_A *Excision*) and to effectively access the definitions, synonyms and semantic types (e.g. *Polypectomy* IS_A *Therapeutic or Preventive Procedure*) of each concept, we followed the methods suggested in [7] and [10], and reorganized the Metathesaurus into four separate tables (named *umls_concept*, *umls_synonyms*, *umls_children* and *umls_sem_types*, see *UMLS* in Fig. 1)

For each question/answer in the questionnaires used in the study and for each diagnostic and relationship, we have looked for one related concept (or several concepts, in the case of complex questions/answers) in the UMLS Metathesaurus.

The identification of mappings has been carried out by six professionals, divided into three teams which have worked in a parallel and independent way. Each team was composed by a technical expert and a medical specialist on colorectal cancer. The technical expert was in charge of helping the medical specialist to navigate through the concepts of the Metathesaurus in order to choose the most appropriate one (or ones) for each element in the questionnaires. This process took 200 hours approximately for each team. The three sets of mappings obtained have been analyzed and compared by the six professionals working together in order to build a final set of correspondences. We selected only those mappings that had been identified (identically or in a very similar way) at least by two of the tree teams.

The final mappings were stored in two separated tables (one for each kind of questionnaire), named *RF_quest_map* and *FA_quest_map* (see *Mappings* in Fig. 1). Each table stores a code associated to each question/answer (*fieldID*) and the identifier of the mapped concept in UMLS (i.e. the Concept Unique Identifier or *CUI*). We used the format *Q_x* to represent the answer *x* of a question *Q*. In the case of complex questions/answers, it was necessary to establish mappings to several concepts in the Metathesaurus. In these correspondences, one concept represents the major part of the semantics, while the others act as complements of the first one. We conceived a unique and innovative representation of this idea, which is based on a numeric column *order*, which indicates the semantic connection of each mapping. We propose that an *order* between 10 and 19 indicates that the concept is a main concept (it represents the major part or all the semantics of the question/answer). Lower levels are



**RF_quest_map**

| fieldID | CUI | order | type |
|---|---|---|---|
| B2a | C0037075 | 10 | Q |
| B2a | C1948054 | 20 | Q |
| B2a | C1279901 | 21 | Q |
| B2a_a | C0439234 | 10 | A |
| B2a_b | C0439673 | 10 | A |
| ... | ... | ... | ... |

**UMLS_concept**

| CUI | prefName | definitions |
|---|---|---|
| C0037075 | Sigmoidoscopy | ... |
| C1948054 | When | ... |
| C1279901 | First | ... |
| C0439234 | Year | ... |
| C0439673 | Unknown | ... |
| ... | ... | ... |

**Fig. 2.** Storage of some complex mappings

associated to stronger semantic connections. An *order* between 20 and 29 indicates that the concept is a complement of the main concept. Higher values of *order* could be used to represent complements of complements. We chose a granularity of 10 for the order variable (i.e. 10-19, 20-29, etc.) because we consider that it is an adequate limit for the maximum number of elements (i.e. main concepts or complements) on the same level. Higher granularities could make the meaning of mappings difficult to understand. The tables of mappings also have a *type* column, indicating if the *fieldID* refers to a question or an answer in the questionnaire (*Q* or *A*).

As an example, in order to map to the UMLS Metathesaurus the question *B2a* (see the last paragraph of section 3.1), we used three concepts from the Metathesaurus: *Sigmoidoscopy*, *When* and *First*. The first one is the main concept (*order* 10), while the other ones act as complements (*order* 20 and 21, respectively). In the DB, this would be stored as three rows in the table *RF_quest_map*, all of them with a *fieldID* equal to *B2a*. The answers to this question are identified with the *fieldID*s *B2a_a* and *B2a_b* (see Fig. 2).

To make this idea human-readable, we propose a representation which is based on showing the complements between brackets. For the previous example, this representation would be: *Sigmoidoscopy (When First)*. This is a possible representation that can be easily understood by humans, and which can be used to develop and intuitive access to the data in future Web query tools.

## 4 Results

We developed methods to map the information from an epidemiological study of colorectal cancer to concepts in the UMLS Metathesaurus, and to store these mappings. We were able to manually map a 92% of the 558 different questions/answers from the questionnaires used in the epidemiological study. We also conceived a text-based representation to express the identified mappings, which will be used to navigate through the collected information.

The results of the mapping process are detailed in Table 2. For the RF and FA questionnaires, the column *Elements* indicates the number of questions in each questionnaire. We have only taken as correct the mappings in which both the question and all its answers were mapped to UMLS.

**Table 2.** Mapping results

| Source | Elements | Mappings |
|---|---|---|
| RF questionnaire | 421 | 387 (92%) |
| FA questionnaire | 126 | 118 (94%) |
| Diagnostics | 5 | 5 (100%) |
| Relationships | 6 | 6 (100%) |
| Total | 558 | 516 (92%) |

## 5 Conclusion

The information collected during an epidemiological study has a great potential of reusability. Annotating these data with ontological terms can considerably facilitate its integration with data from other studies, allowing the development of new studies over a larger volume of data.

This paper presents innovative methods to annotate the information gathered during an epidemiological study of colorectal cancer with terms from a well-known set of medical ontologies, the UMLS Metathesaurus. This annotation may help to that the collected information be representation-independent, in such a way that can be understood from other parts of the world where they use languages or terminologies different from ones in the region where this study was achieved, facilitating the integration of these data with data from other information resources.

At the moment, we are developing a prototype of Web interface which use the established mappings to allow browsing and querying the gathered data in text mode (using the brackets-based representation mentioned in section 3.2). Moreover, we are studying the development of Web services [11] to allow that other information systems can remotely access these data.

**Acknowledgments.** This work was partially supported by the following: Spanish Ministry of Education and Science (Ref TIN2006-13274); European Regional Development Funds; grant (Ref. PIO52048) from by the Carlos III Health Institute; grant (Ref. PGIDIT 05 SIN 10501PR) from the General Directorate of Research of the Xunta de Galicia; and grant (File 2006/60) from the General Directorate of Scientific and Technologic Promotion of the Galician University System of the Xunta de Galicia. The work of José M. Vázquez is supported by an FPU grant (Ref. AP2005-1415) from the Spanish Ministry of Education and Science.

## References

1. Bodenreider, O.: The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Research 32, 267–270 (2004)
2. Vazquez, J.M., Martinez, M., Lopez, M.G., Gonzalez-Conde, B., Arnal, F.M., Pereira, J., Pazos, A.: Development of an Information System for Multicenter Epidemiologic Studies on Cancer. In: Information Technology Applications in Biomedicine, Tokio, Japan, 2007. ITAB 2007, pp. 149–152 (2007)

3. Pérez-Rey, D., Maojo, V., García-Remesal, M., Alonso-Calvo, R., Billhardt, H., Martin-Sánchez, F., Sousa, A.: ONTOFUSION: Ontology-based integration of genomic and clinical databases. Comput. Biol. Med. 36, 712–730 (2006)
4. Barrasa, J., Corcho, O., Gómez-Pérez, A.: R2O, an Extensible and Semantically Based Database-to-Ontology Mapping Language. In: SWDB 2004 (2004)
5. Konstantinou, N., Spanos, D., Chalas, M., Solidakis, E., Mitrou, N.: VisAVis: An Approach to an Intermediate Layer between Ontologies and Relational Database Contents. In: Proc. Int. Workshop on Web Information Systems Modeling (2006)
6. Xu, Z., Zhang, S., Dong, Y.: Mapping between Relational Database Schema and OWL Ontology for Deep Annotation. In: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence, pp. 548–552 (2006)
7. Shah, N.H., Rubin, D.L., Supekar, K.S., Musen, M.A.: Ontology-based Annotation and Query of Tissue Microarray Data. In: AMIA Annu. Symp. Proc., vol. 709, p. 13 (2006)
8. Wache, H., Vögele, T., Visser, U., Stuckenschmidt, H., Schuster, G., Neumann, H., Hübner, S.: Ontology-based integration of information-a survey of existing approaches. In: IJCAI-01 Workshop, pp. 108–117 (2001)
9. de Coronado, S., Haber, M.W., Sioutos, N., Tuttle, M.S., Wright, L.W.: NCI Thesaurus: Using Science-based Terminology to Integrate Cancer Research Results. Med. info. 11, 33–37 (2004)
10. Pisanelli, D.M., Gangemi, A., Steve, G.: An Ontological Analysis of the UMLS Metathesaurus. In: Proc. AMIA Symp., vol. 810, p. 4 (1998)
11. Maojo, V., Crespo, J., de la Calle, G., Barreiro, J., Garcia-Remesal, M.: Using Web Services for Linking Genomic Data to Medical Information Systems. Methods of Information in Medicine 46, 484–492 (2007)

# A Primer in Knowledge Management for Nanoinformatics in Medicine

Fernando Martín-Sanchez, Victoria López-Alonso, Isabel Hermosilla-Gimeno,
and Guillermo Lopez-Campos

Medical Bioinformatics Dept., Institute of Health Carlos III,
Ctra. Majadahonda-Pozuelo Km2., 28220, Majadahonda, Spain
{fms, victorialopez, isahermo, glopez}@isciii.es

**Abstract.** In the last years new scientific knowledge and technological developments derived from the Human Genome Project have been affecting the way in which biomedical research and clinical practice is carried out. This fact has had a profound impact in the current scope and definition of the"biomedical informatics" discipline. The big "genomic wave" boosted efforts in the medical informatics community to adapt its systems and make it possible the integration of "omics" knowledge and bioinformatic tools. Recent advances in another discipline, Nanotechnology, seem to be similarly starting to affect clinical research and practice (nanomedicine). The purpose of this document is to review the state of the art and current developments in nanomedicine and its potential impact on Biomedical Informatics. The authors have recently been granted with an European Commission research project, Action-GRID, that will analyse the field of nanoinformatics with the aim of identifing needs and discuss ongoing challenges and priorities in terms of knowledge management and discovery in nanomedicine.

**Keywords:** nanomedicine, nanoinformatics, knowledge management.

## 1 Introduction

In recent years nanotechnology has developed greatly designing elements for diagnosis and therapy in the biomedical area. At it has been pointed out, it is important to place emphasis in three critical research areas, namely tools, informatics and nanoscale building to further advance in this discipline [1]. The large amount of information generated creates a challenge to those that need to manage it, this is to organize, standardize, store, compare, analyse and visualise all the information to extract useful and applicable conclusions. Nanoinformatics arises as the discipline that can manage this information overload and help draw useful applicable results [2]. This article analyses the impact that this new discipline may have in biomedical informatics and provides a general framework for knowledge management in nanomedicine.

The main nanotechnologies developed to date are: a) image processing where nanoparticles act as contract agents for seeing cells and tissues at a finer scale,

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 66–72, 2008.

b) modelling and simulation for testing hypothesis and predicting behaviour, c) Nanolibrary [1] with interconnected databases and tools, virtual labs called collaboratories where data and instruments are shared remotely, d) databases for nanomaterials such as the nanoparticle Information Library (NIOSH) or the Nanocharacterization lab (NCI) and, e) linking nanodatabases with other important existing resources in other fields.

The advancement in these areas will result in the development of a new and improved generation of medical devices and treatments. In research they will provide insight about how diseases develop. As an example, next generation DNA sequencing relies on nanopore based technology.

In the diagnostic field biomarkers, nanoparticles and imaging can be used as high resolution agents for tissue imaging, for instance with tumours. They will also be used for continuous monitoring of biochemical entities and other blood parameters.

In the therapeutic field, novel nanostructures will be used as drugs against cancer, neurodegenerative or cardiovascular disease. Some examples are AMGEN´s Neulasta [3, 4] or Abraxane [5], a drug for treating metastatic breast cancer that was approved by the FDA in 2005. This is an example of a new delivery method of a chemotherapy drug. It consists of paclitaxel bound to nanoparticles of albumine. At this time there are more than 100 products being reviewed. There are also new chemotherapy delivery systems such as buckyballs, nanocapsules and dendrimers. Many drugs are being designed from the bottom-up; they are artificial nanostructures that mimic complex biomolecules and that show regenerative power for neurodegenerative disease. Finally a big advancement will be seen in regenerative medicine through the manufacture of new kinds of artificial tissues to replace kidney, liver or to repair nerve damage, or using nanotubes and nanofibers to build scaffolds where to grow networks of cells from different organs.

## 2 Biomedical Informatics and Nanotechnology

It is probable that just as genomics brought about the production of a plethora of new data and the development of bioinformatics, it is safe to assume that the integration of nanotechnologies will lead advancements beyond genomic medicine with big impacts in several areas. Advancement in prediction of diseases will allow for early diagnosis analysing hundreds of substances in the blood and estimating disease risks. There will also be a pre-emptive effect making possible the early detection of diseases such as diabetes permitting a rapid treatment protocol. The impact will also be seen in personalised medicine where it will be feasible to detect variations in disease state and to tailor therapies, adjusting drug cocktail to the individual patient.

All the new advances in nanotechnology are encouraging the scientific community, especially in biomedicine, to use the ‘nanoscope’ that will allow them to see a new scale of genomic, phenotypic and environmental data to integrate with the information already available (Figure 1). Nanosensors and nanodevices will permit to assess the different positive and negative factors that affect an individual´s risk to have

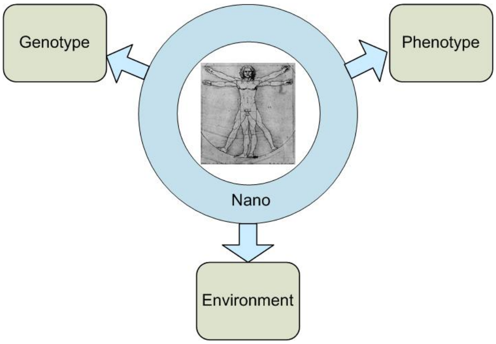


**Fig. 1.** The `nanoscope´ vision in biomedicine

a disease. This new approach demands collaboration between many different disciplines that deal with physics, chemistry, biology, engineering, information technology, and medicine.

Over the years information technology and computer science have been developing new research disciplines at the same time that the new trends in biomedical science, medicine and health care advanced. Figure 2 shows a spectrum of the informatics disciplines developed over different application domains. Public Health Informatics, addresses information arising from populations, Medical Informatics focuses on the individual and Imaging Informatics works with tissues and organs while Bioinformatics deals with molecular and cellular processes. Thus, as illustrated in the diagram, informatics has important contributions to make across that entire data spectrum and levels of biocomplexity, with the development of chemo-informatics, neuroinformatics and in the last years with the development of nanoinformatics [6].

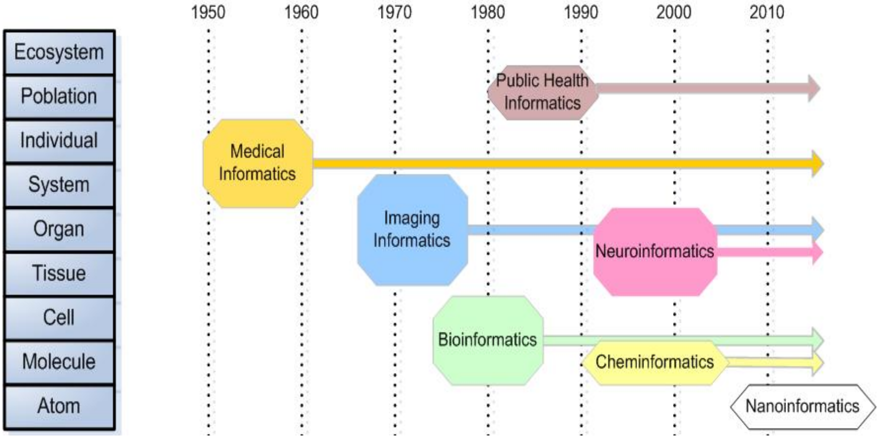


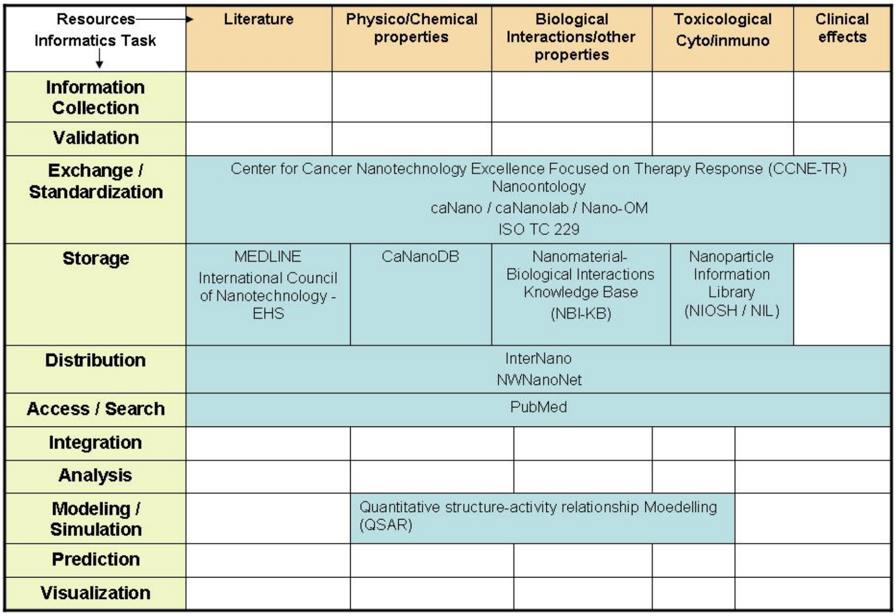
**Fig. 2.** Spectrum of the informatics disciplines across time (as one moves from left to right) and over different levels of biocomplexity from atom to ecosystem (bottom-up) domains

## 3 The European Project Action-Grid: Towards a Knowledge Management Framework for Nanomedicine

With this perspective in mind a group of researchers from 6 countries got together to present a proposal of a project to the European Commission (EC) which funded it in 2008. ActionGrid analyzes synergies between Medical Informatics (MI), Bioinformatics (BI) and Nanoinformatics (NI). The project has as its main objective to act as a multiplier of previous outcomes in GRID and Biomedical Informatics (BMI) and to combine these results with data from an inventory of Grid/Nano/BMI methods and services developed by the consortium. These outcomes will be also disseminated in Latin America, the Western Balkans and Northern Africa. The obtained results will be presented in a White Paper written in collaboration with a panel of recognized experts in various fields. This document will be delivered to the EC to establish a future agenda covering the Grid/Nano/Bio/Medical Informatics areas and to develop new plans in Latin America, the Western Balkans and Northern Africa.

The main impact expected is to expand previous initiatives to create a common health information infrastructure in Europe, and extending it to other regions. It is expected that it will enhance cooperation between research centres, universities, hospitals, and industry. ACTION-Grid will expand the impact of EC achievements in Grid and BMI to researchers, educators, and health practitioners' world-wide [7].

In next years hundreds of nanotechnology developments, including drugs and medical devices should be processed as basic information. Nanotechnology requires



| Resources → <br> Informatics Task ↓ | Literature | Physico/Chemical properties | Biological Interactions/other properties | Toxicological Cyto/inmuno | Clinical effects |
|---|---|---|---|---|---|
| Information Collection | | | | | |
| Validation | | | | | |
| Exchange / Standardization | Center for Cancer Nanotechnology Excellence Focused on Therapy Response (CCNE-TR)<br>Nanoontology<br>caNano / caNanolab / Nano-OM<br>ISO TC 229 | | | | |
| Storage | MEDLINE<br>International Council of Nanotechnology - EHS | CaNanoDB | Nanomaterial-Biological Interactions Knowledge Base (NBI-KB) | Nanoparticle Information Library (NIOSH / NIL) | |
| Distribution | InterNano<br>NWNanoNet | | | | |
| Access / Search | PubMed | | | | |
| Integration | | | | | |
| Analysis | | | | | |
| Modeling / Simulation | | Quantitative structure-activity relationship Moedelling (QSAR) | | | |
| Prediction | | | | | |
| Visualization | | | | | |

**Fig. 3.** Summary of tools that are being developed for nanomedicine knowledge management

informatics tools to process all the knowledge arising from the nanoworld and to integrate it with the biomedical (phenotypic, genotypic and environmental) domain. Nanomedicine gathers and deals with large volumes of complex data, linked with external sources and usually distributed in heterogeneous locations. However, there are some differences with other biomedical data, like genomic ones that are predictive and personal particularly related to the nature of nanotechnology and its exogenous origin.

Nanoinformatics involves the development of effective tools/technologies for collecting, standardizing, sharing, analyzing and visualizing information relevant to the nanoscale science in several areas such as literature, physico-chemical properties, biological and toxicological interactions and clinical effects.

A brief overview of the tools that are being developed to manage and process nanomedicine knowledge is detailed in the following paragraphs and they are summarized in Figure 3.

## 4 Standardisation Initiatives for Nanomedicine Knowledge Representation

Development of standards and ontologies in nanotechnology is one of the top priorities as it allows integration with other data. Nanotechnology standards should be linked or integrated with some common biomedical controlled vocabularies such as Mesh, SNOMED, LOINC, CPTs, DICOM, CheBi and PubChem, that deal with literature, medical procedures, laboratory testing, images, or chemicals. Other ontologies will probably not be greatly affected by the new data. GO and ICD would fall in this category, provided that the definition of genes or diseases will not be much affected by the new nano procedures and techniques..

The National Cancer Institute Alliance for Nanotechnology is working in systems and methods for data standardization and classification to facilitate and improve knowledge management, examples of these are the following initiatives:

- The Centers for Cancer Nanotechnology Excellence [8] that are developing a common strategy for sharing data across disciplines focused on therapy response [8] structuring data to make it caBIG compatible and providing informatics resources for managing data and experiments.
- The Nanotechnology Characterization Laboratory [9] is testing pre-clinical toxicology, pharmacology, and efficacy of nanoparticles and devices using as interface caNanoLab. caNanoLab is based on the nanoparticle characterization object model (Nano-Om), an initial standard representation of nanoparticle characterizations.

A big effort is placed also in the development of the ISO/TC229 standardization system in the field of nanotechnologies in three categories: -terminology and nomenclature, -measurement and characterization and -health, safety and environmental standards [10].

## 5 Nanomedicine Knowledgebases

Scientific literature about Nanotechnology and Nanomedicine can be found at general biomedical databases, such as MEDLINE, and in other resources, such us the International Council of Nanotechnology Environmental Health Safety (ICON EHS) that includes more specific “nano” information.

CaNanoDB [11] is developing a nanoparticle informatics resource that includes a database of all available nanoparticle technologies and a toolbox for modeling of targeted drug delivery and diagnostics using this novel technology.

Researchers at Oregon State University and the Oregon Nanoscience and Microtechnologies Institute (ONAMI) [12] are developing a collaborative knowledgebase, the Nanomaterial Biological Interactions (NBI) Knowledgebase for annotated data on nanomaterial characterization. It includes chemical and physical properties, synthesis methods, and nanomaterial-biological interactions.

The Nanoparticle Institute for Occupational Safety and Health (NIOSH) is developing a web-based Nanoparticle Information Library (NIL) to share information on nanomaterials, including their health and safety-associated properties [13].

## 6 Knowledge Exchange

The National Nanomanufacturing Network (NNN) is developing InterNano, an open-source, online information clearinghouse and digital library to aggregate, organize, standardize, and disseminate nanomanufacturing information. It encourages the free exchange of information, and supports the ways that researchers have come accustomed to gathering information using today's web-based tools.

In this respect ONAMI is continuing to invest in the NWNanoNet™ [14], a network of shared user facilities that provide advanced measurement and fabrication services to academic and industry researchers.

## 7 Nanostructures Modeling and Simulation

Quantitative structure-activity relationship (QSAR) is the process by which the chemical structure is quantitatively correlated with a well defined process, such as biological activity or chemical reactivity. The development of Quantitative Structure-Activity Relationships (QSARs) requires

- Well characterized material structures,
- Experimental data establishing the activity of those structures in relevant environments, and
- A sound theoretical basis for the mechanisms underlying the relationship between them.

## 8 Conclusion

The Nanoinformatics tools list is still relatively empty, filling it with more powerful instruments would translate into gains in scientifically useful knowledge, the starting point for next-generation nanomedicine.

As this field advances, biomedical informatics will be posed again with new challenges related to the management and discovery of knowledge that can be used for decision making at the point-of-care in clinical routine along the coming years. As the "-omic" revolution has been influencing the research roadmap in biomedicine, nanomedicine and regenerative medicine open new avenues for knowledge representation and biomedical informatics methods and tools.

In this regard, education is very important as a new generation of scientists familiar with all the data generated from nanotechnology will be essential to manage this new knowledge and adequately integrate it with existing and developing biomedical data to best translate all these findings into better healthcare.

## References

1. Schmidt, K.: Nanofrontiers: Visions for the future of Nanotechnology. Woodrow Wilson
2. Ruping, K., Sherman, B.W.: Nanoinformatics: Emerging Computational Tools in Nanoscale Research. In: Nanotech 2004: Technical Proceedings of the 2004 NSTI Nanotechnology Conference and Trade Show, vol. 3 (2004)
3. National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov
4. Moran, N.: Nanomedicine lacks recognition in Europe. Nature Biotech. 4(2), 121 (2006)
5. Abraxane website, http://abraxane.com
6. NNN – National Nanomanufacturing network. NSF (07) In: Workshop on Nanoinformatics Strategies, Arlington, VA (June 2007), http://128.119.56.118/~nnn01/Workshop.html
7. ActionGrid Proposal (not published)
8. http://med.stanford.edu/rmg/funding/CCNE_TR.html
9. http://ncl.cancer.gov/
10. ISO TC 229 Webpage, http://www.iso.org/iso/iso_technical_committee.html?commid=381983
11. http://gforge.nci.nih.gov/projects/canano/
12. http://www.onami.us/
13. http://www2a.cdc.gov/niosh-nil/
14. http://nnn.internano.org/
15. http://www.onami.us/NanoNet/

# Towards Natural Head Movement of Autonomous Speaker Agent

Marko Brkic[1], Karlo Smid[2], Tomislav Pejsa[1], and Igor S. Pandzic[1]

[1] Faculty of electrical engineering and computing, Zagreb University, Unska 3,
HR-10000 Zagreb, Croatia
{marko.brkic, tomislav.pejsa, igor.pandzic}@fer.hr
[2] Ericsson Nikola Tesla, Krapinska 45, p.p. 93, HR-10 002 Zagreb
karlo.smid@ericsson.com

**Abstract.** Autonomous Speaker Agent (ASA) is a graphically embodied animated agent capable of reading plain English text and rendering it in a form of speech, accompanied by appropriate, natural-looking facial gestures [1]. This paper is focused on improving ASA's head movement trajectories in order to achieve facial gestures that look as natural as possible. Based on the gathered data we proposed mathematical functions that, using two input parameters (maximum amplitude and duration of the gesture) generate natural-looking head motion trajectory. Proposed functions were implemented in our existing ASA platform and we compared them with our previous head movement models. Our results were shown to a larger number of people. The audience noticed that results showed improvement in head motion and didn't detect any patterns which would suggest that animation was done with predefined motion trajectories.

## 1 Introduction

While animated motion pictures such as Shrek feature truly impressive and natural-looking animation of character's faces, they require tremendous effort by highly skilled and talented artists. Production of facial animations in a fully automated way, without any human intervention, while striving to reach similar levels of fidelity, is a very demanding process. In Section 2 we give a brief overview of research efforts in this field including our own research on the system we call the Autonomous Speaker Agent (ASA).

ASA (Fig. 1 shows snapshots of an ASA while talking) is an extension of a Visual Text-to-Speech (VTTS) system. A classical VTTS system produces lip movements synchronized with the synthesized speech based on the timed phonemes generated by the speech synthesis. Since VTTS system can only obtain phonetic information from the speech synthesis, it has no basis for generating realistic gestures. This problem is solved by an approach that combines the lexical analysis of input text with a statistical model describing the dynamics, frequencies and amplitudes of facial gestures [1].

ASA is based on HUman GEsturing (HUGE) [2] software architecture for production and use of statistical models for facial gestures based on inducement of arbitrary kind. An inducement is a signal that occurs in parallel to the production of facial gestures in

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 73–80, 2008.

human behaviour and that may have a statistical correlation with the occurrence of facial gestures, e.g. text that is spoken, audio signal of speech, bio signals, emotions etc. The correlation between the inducement signal and facial gestures is used to first build the statistical model of facial gestures based on a training corpus consisting of sequences of gestures and corresponding inducement data sequences. In the runtime phase, the raw, previously unknown inducement data is used to trigger (induce) the real time facial gestures of the agent based on the previously constructed statistical model.

**Fig. 1.** Autonomous Speaker Agent

In first version of ASA, head motion animation was implemented using sine function trajectory. This paper expands on this research by trying to improve head motion of facial gestures in order to get more natural-looking facial gestures.

In order to capture the dynamics of real gestures, we analyzed recorded video clips of real professional TV speakers as a starting point for extraction of head movement coordinates. Section 3 explains which facial gestures were analyzed and how measurements were done in order to obtain head movement trajectories.

The following two sections deal with proposing the facial gestures animation functions by fitting a mathematical model to the recorded motion trajectories. Section 4 explains the analysis of data gathered from nod gesture motion analysis and proposes a way to interpret this data using trigonometric functions. In a similar way Section 5 explains swing motion and proposes a way to interpret this gesture.

Section 6 elaborates achievements we made in this project and we conclude the paper with the conclusion and a discussion of future work.

## 2 Background

State of the art research on real-time natural facial gesture animation is very intensive field. Busso et al. [3] present a novel data-driven approach to synthesize natural human head motions driven by speech prosodic features. The problem was modeled as classification of discrete representations of head poses. Hidden Markov Models (HMM) [4] were used to learn the temporal relation between the dynamics of head motion sequences and the prosodic features. Using new speech material, the HMM works as a sequence generator for the most likely head motion sequences. In the end bi-grams and spherical cubic interpolation techniques are used to smooth the synthesized sequence. Hofer et al. [5] models longer units of motion and speech and to reproduce their trajectories during synthesis, they utilize a promising time series stochastic model called “Trajectory Hidden Markov Models” [6]. The BEAT system [7] uses linguistic and contextual information contained in a text to control the movements of hands, arms and a face, and the intonation of a voice. The mapping from a text to the facial, intonational and body gestures is contained in a set of rules derived from a state of the art research in nonverbal conversational behavior. That mapping also depends on the knowledge base of an Embodied Conversational Agent ECA environment. That knowledge base is populated by a user who animates the ECA so the production of the facial gestures is not automatic. Furthermore, the system only provides output for animation systems but it does not introduce any animation model for ECA facial gestures. Also, in the current set of supported nonverbal behaviors, head nods and eyes blinks must be included.

In our approach facial gesture animation is driven by linguistic and contextual structure of uttered English text. The goal is to enable the ASA to perform gestures that are not only dynamically correct, but also correspond to the underlying text [1]. Input for animation models are facial gesture maximal amplitude and duration. Each gesture trajectory was nose tip movement that needed to be represented with some function (Fig. 2).

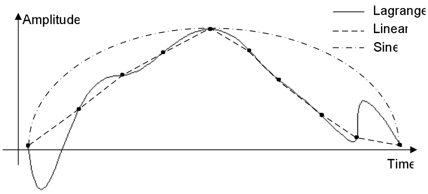


**Fig. 2.** Linear and Lagrange implementation

First approach is to connect neighboring coordinates with linear function so each gesture has its own subintervals (time frames) defined with those linear functions. If ASA animation module requests amplitude calculation for an exact time point, we have to call the linear function the time frame of which includes this time point.

Second approach is to use Lagrange interpolation. Theory says that a set of N points can be interpolated by a polynomial of degree at most N-1 [8]. If we had ten coordinates for one gesture (Fig. 2.) after using interpolation we would get one function of degree (at most) nine. Resulting polynomial is:

$$P(x) = \sum_{j=1}^{n} \left( y_j \prod_{\substack{k=1 \\ k \neq j}}^{n} \frac{x - x_k}{x_j - x_k} \right) \qquad (1)$$

Those two approaches aren't good enough for several reasons: linear implementation has sharp edges and observer can notice that virtual character has discontinued movement at interval boundaries; Lagrange implementation has sudden large amplitude movement at the beginning and at the end of intervals; there is only one time interval for each gesture (with constant length and maximal amplitude) and after observing ASA after some time it is obvious that head always moves in the same way.

Trigonometry sine implementation solves the problem regarding trajectory constant length and amplitude. However, implementation still doesn't solve the problem of describing natural-looking gestures since we know that sine function is not an ideal representation of facial motion (refer to Fig 2.). This paper explains how sine implementation was improved.

## 3 Head Movement Analysis

Within head movement we distinguish two types of facial gestures: nods and swings. Nod is an abrupt swing of the head with a similarly abrupt motion back and swing is an abrupt swing of the head without the back motion with following directions: up, down, left, right and diagonal (divided into four quadrants). We have following nod directions: up and down, down and up, left and right, right and left and diagonal (divided into four quadrants). The parameters that are important for head movement are maximum amplitude and duration. To get these parameters we analyzed 56 video clips originating from Ericsson's “5minutes” web cast. To describe those videos we used HUGE [2] gesture XML files containing gesture type, maximum amplitude, start and end time for every identified head movement. The best way of tracking head movement is by observing the speaker’s nose tip position as a reference point. Obtained coordinates were used to get graphical representation of head movement trajectory. In order to ensure that all measurements taken from these videos are in the same proportion, we normalized them using the Mouth-Nose Separation unit (MNS0): a distance between (in pixels) nose tip and upper lip. Excel was used to create graphs from obtained data and those graphs were used for gesture motion analysis. In the next chapter we elaborate on nod head movements.

## 4 Synthesizing Nod Gestures

In this chapter we explain graphs of gathered data for nod gesture and propose a function that describes that specific gesture trajectory. After all measurements were done and everything was placed in tables, graphs were made using Microsoft Excel for each gesture. Since all measurements in nod gestures show same properties, graph of nod up gesture is enough to comment all nod gestures. The only difference between all nod motions is in motion direction.

Functions drawn in graphs using obtained data were used for gesture motion analysis. After graph analysis it was concluded that trigonometry functions were the best solution for describing nod trajectory functions. Since most of the functions on graph were not exact sine functions some statistic analysis needed to be done. After analysis we divided our proposed functions in three types where the first one is a function where its time of maximum amplitude value is to the left of the function duration midpoint (D/2), the second one is ideal sine function (time of maximum amplitude is exactly in the middle) and the third one with maximum positioned to the right of the duration midpoint.

First and third functions are divided in two intervals where time of maximum is the separation point. Fig. 3. shows all three functions with characteristic parameters.

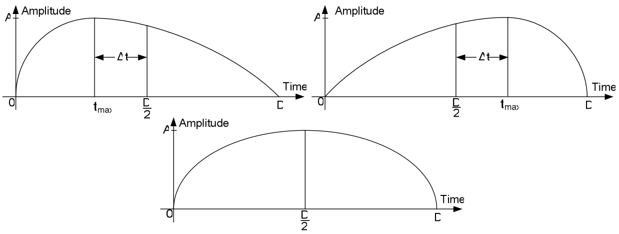


**Fig. 3.** Three characteristic functions of nod gesture

Sine function is used to describe the ascending part of the function, and cosine function is used to describe the descending part of function. In order to calculate percentages by which three different functions appear we analyzed all data from obtained functions and compared maximum amplitudes of each function with amplitude value from half of the interval of the same function. Percentages for nod up are given in table 3.

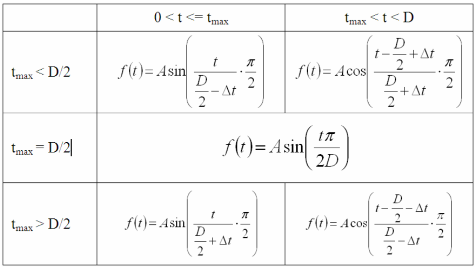
During data analysis we couldn't find any connection between Δt value (a difference in milliseconds between function maximum and functions half-interval value) and facial gesture. We only concluded that Δt value can be inside [0, D/4] interval. This interval was used to randomly generate Δt value for each head motion.

**Table 1.** Half interval deviation percentage values for nod up

| | $t_{max}$ < D/2 | $t_{max}$ = D/2 | $t_{max}$ > D/2 |
|---|---|---|---|
| Nod up | 41% | 11% | 48% |

Function definitions shown in Table 4 are same for all nod gestures. With these three functions we tried to describe the behavior of node gesture. By implementing Δt we managed to describe different velocities of head movement in two different directions, one from the start position to the maximum and second one in the opposite direction. The next chapter explains swing gesture implementation.

**Table 2.** Function definitions for nod gesture



| | $0 < t <= t_{max}$ | $t_{max} < t < D$ |
|---|---|---|
| $t_{max}$ < D/2 | $f(t)=A\sin\left(\frac{t}{\frac{D}{2}-\Delta t}\cdot\frac{\pi}{2}\right)$ | $f(t)=A\cos\left(\frac{t-\frac{D}{2}+\Delta t}{\frac{D}{2}+\Delta t}\cdot\frac{\pi}{2}\right)$ |
| $t_{max}$ = D/2\| | $f(t)=A\sin\left(\frac{t\pi}{2D}\right)$ | |
| $t_{max}$ > D/2 | $f(t)=A\sin\left(\frac{t}{\frac{D}{2}+\Delta t}\cdot\frac{\pi}{2}\right)$ | $f(t)=A\cos\left(\frac{t-\frac{D}{2}-\Delta t}{\frac{D}{2}-\Delta t}\cdot\frac{\pi}{2}\right)$ |

## 5 Synthesizing Swing Gestures

This chapter explains graphical results of swing gesture measurements and proposes a function that describes its motion trajectory. Swing gesture is a head movement

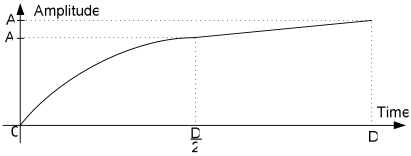


**Fig. 4.** Two characteristic functions of swing gesture

without return motion to starting position (like in nod gestures). Previous sine implementation used first quarter of sine period for reaching maximum coordinate, but analysis done for this movement showed that after the first half of swing gesture duration maximum coordinate is reached with linear trajectory (Fig 4.).

Based on graph in Fig. 4. all swing functions are defined in Table 5:

**Table 3.** Function definitions for swing gesture

| | |
|---|---|
| 0 < t <= D/2 | $f(t) = A\sin\left(\frac{t\pi}{2D}\right)$ |
| D/2 < t < D | $f(t) = \frac{2t(A - A')}{D} + 2A' - A,\ A' = A\sin\left(\frac{\pi}{4}\right)$ |

By dividing the function into two segments we tried to simulate the increase of head movement velocity in the second part of function duration.

# 6 Results

To determine the effect of new head gestures on naturalness of virtual characters' motion, we conducted a subjective test. We generated two animation sequences, one with old sine-based head movement and another with new trigonometric gestures. Both sequences featured the same virtual character presenting the same text and making the same facial and head gestures. The sole difference between the sequences was the mathematical model used for head movement.

The test sequences were shown to groups of subjects. There was a total 185 subjects. All of them are students in fields of Computer Science and Telecommunications and none of them are familiar with our area of research. After viewing a test sequence, the subjects were asked to score the virtual character by answering the following questions:

1. Did the character's head movements appear natural (5) or not (1)?
2. Did the character on the screen appear interested in (5) or indifferent (1) to you?
3. Did the character appear engaged (5) or distracted (1) during the conversation?
4. Did the personality of the character look friendly (5) or not (1)?

Note that possible scores are 1 to 5, where higher scores correspond to more positive attributes in the speaker.

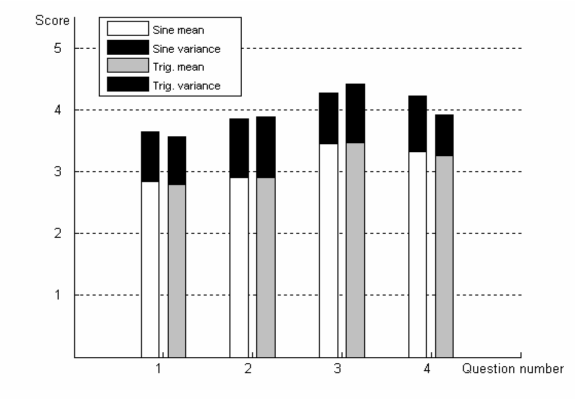


**Fig. 5.** Test scores for virtual characters with old and new head movements

Our testing has shown that new head movements have no measurable effect on subjects' subjective perception of the virtual character. These findings were confirmed when we performed one-way ANOVA on the results and determined that virtual characters in the two test sequences did not receive significantly different scores on any of the 4 questions.

However, it must be noted that a small number of subjects were explicitly asked to compare head gestures in the two animation sequences and they invariably noted an improvement in naturalness for trigonometric head motion.

## 7 Conclusion and Future Work

ASA combines lexical analysis of English text and statistical models of facial gestures in order to generate the gestures related to the spoken text. Current state of the art work in the field of head movement animation trajectories is mostly based in correlating head movement with the prosody features of the speech. In our approach we use the correlation between the lexical structure of spoken text and head motion. Also, we introduce rather simple mathematical techniques as a base for animation model. Our goal is to test practical value of computationally low intensive models. Progress explained in this paper should be just a beginning in improving head gestures of virtual characters. We still have to analyze diagonal head motion and eyebrows motion in order to improve the basic sine trajectory model. Future work should include four main diagonal implementations both in nod and swing head motions and eyebrow raise and frown motion. Goal of this paper is to improve animation of virtual speakers in order to achieve better automatic generation of head gestures for applications such as newscasters and storytellers. We improved nod and swing motion and gave a solid base ground for further improvement and research. This future research should be based on a larger data corpus.

## Acknowledgment

This work was partly carried out within the research project "Embodied Conversational Agents as interface for networked and mobile services" supported by the Ministry of

Science, Education and Sports of the Republic of Croatia and within the research project "Summer Camp 2007" supported by Ericsson Nikola Tesla.

## References

1. Smid, K., Pandzic, I.S., Radman, V.: Autonomous Speaker Agent. In: Proceedings of the Computer Animation and Social Agents Conference CASA 2004, Geneva, Switzerland (2004)
2. Smid, K., Zoric, G., Pandzic, I.S.: [HUGE] Universal Architecture for Statistically Based HUman GEsturing. In: Gratch, J., Young, M., Aylett, R.S., Ballin, D., Olivier, P. (eds.) IVA 2006. LNCS (LNAI), vol. 4133, pp. 256–269. Springer, Heidelberg (2006)
3. Busso, C., Deng, Z., Neumann, U., Narayanan, S.: Natural head motion synthesis driven by acoustic prosodic features. Computer Animation and Virtual Worlds 16(3-4), 283–290 (2005)
4. Rabiner, L.R.: A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proceedings of the IEEE 77(2), 257–286 (1989)
5. Hofer, G., Shimodaira, H., Yamagishi, J.: Speech driven head motion synthesis based on a trajectory model. In: International Conference on Computer Graphics and Interactive Techniques. ACM SIGGRAPH 2007 posters, article no. 86, San Diego, USA (2007)
6. Zen, H., Tokudaa, K., Kitamura, T.: Reformulating the HMM as a trajectory model by imposing explicit relationships between static and dynamic feature vector sequences. Computer Speech & Language 21(1), 153–173 (2007)
7. Cassell, J., Vilhjálmsson, H., Bickmore, T.: BEAT: the Behavior Expression Animation Toolkit. In: SIGGRAPH 2001, pp. 477–486. ACM, New York (2001)
8. Abramowitz, M., Stegun, I.A. (eds.): Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, 9th printing, pp. 878–879, 883. Dover, New York (1972)

# Automatic Translation in Two Phases: Recognition and Interpretation[*]

Osvaldo Cairó and Silvia Guardati

Department of Computer Science
Instituto Tecnológico Autónomo de México (ITAM)
Río Hondo 1, 01080 México DF, México
{cairo, guardati}@itam.mx

**Abstract.** This paper presents an approach to automatic translation in two steps. The first one is recognition (syntactic) and the last one is interpretation (semantic). In the recognition phase, we use volatile grammars. They represent an innovation to logic-based grammars. For the interpretation phase, we use componential analysis and the laws of integrity. Componential analysis defines the sense of the lexical parts. The rules of integrity, on the other hand, are in charge of refining, improving and optimizing translation. We applied this framework for general analysis of romance languages and for the automatic translation of texts from Spanish into other neo-Latin languages and vice versa.

**Keywords:** Natural language, automatic translation, volatile grammars.

## 1 Introduction

Formalization and translation of natural languages has been a topic of great interest among linguistic and computer researchers. But its complexity and richness has kept it as an open problem.

Natural language is the mean used by man to communicate, to establish political and business relationships among peoples, and to express ideas and emotions. The history of machine translation, on the other hand, begins in the 1950s, after World War II. The Georgetown experiment (1954) involved fully-automatic translation of over sixty Russian sentences into English. The experiment was a great success and the authors claimed that within three to five years, machine translation would be a solved problem. However, several problems regarding the syntactic and semantic analysis of the sentence have been found, additionally to the ambiguities, so rich and diverse in natural languages. During the past few years, particularly, significant advancements have been achieved, both in the processing of natural language ([1], [2]), and in machine translation ([3], [4], [5], [6]). *Systram*, *BabelFish*, *PROMPT*, *Softissimo Reverso*, *LinguaSoft*, and *LinguaWeb* represent a proof of these advances. However, current systems are unable to produce output of the same quality as a human translator, particularly where the text to be translated uses casual language. The reason is clear: many problems remain the same.

[*] This project has been partially funded by Asociación Mexicana de Cultura A.C.

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 81–88, 2008.

The purpose of research on this subject is to solve some of these problems. It focuses on the general analysis of romance languages and, mainly, in the development of tools to perform automatic translation of texts from Spanish into other romance languages and *vice versa*. We are interested in the *romance* languages because they have similar syntax, vocabularies and morphological structure, although as it can be supposed they have chronological, geographical and historical differences. These are the causes that have definitely segregated and channeled, into different currents, the languages of the ancient provinces of the Roman Empire; segregation that lead to the birth of romance languages and to the formation of Spanish.

Automatic translation of texts is achieved by means of a bi-directional translator, that can translate sentences from an *A* language into sentences of a *B* language or vice versa. A bi-directional translator is very helpful for the translation of texts between different dialects that belong to a single language, or between languages that belong to a same parent language, for example romance languages that come from Latin.

## 2 Using Volatile Grammars in the Recognition Phase

From the first formalism presented by Colmeraurer about grammars in logic, *metamorphosis grammars* [7], several modifications or innovations have been presented: defined clause grammars, DCGs [8], extraposition grammars, XGs [9], discontinued grammars [10], reversible grammars [11], evolving grammars [12] and even string variable grammars [13], a modification of DCGs, for DNA sequences. *Volatile grammars* (VG) represent an innovation to logic-based grammars.

Given an entry sentence in natural languages, volatile grammars obtain the structural description of the sentence through the automatic generation of a syntagmatic tree. The central idea of these new grammars consists in having only basic rules and generating rules. In other words, rules that allow dynamically generating new rules, if necessary, but that disappear after being used saving space. Hence, the name *volatile* given to these grammars. In recent publications about computational linguistics, economy in the number of rules has been repeatedly mentioned as one of the main objectives. Volatile grammars emphasize, on this purpose: minimization of the number of rules of a grammar.

We can formally define *VGs* with the following quadruplet: $VG = \{\Sigma, T, N, P\}$, where $\Sigma$ represents the initial symbol, $T$ the terminal symbols, $N$ the non-terminal symbols and $P$ the productions.

- $\Sigma$ is always a symbol in N.
- $T$, $N$ and $P$ represent finite sets. $T \cap N = \varnothing$
- We conventionally denote $T \cup N$ with $A$ (alphabet).
- The set of productions $P$ consists of expressions in the form:

  $\alpha \rightarrow \beta$ or $\alpha \rightarrow GENERATES\ (S_1, S_2, ..., S_n)$

  where the generating rule has the form:

  $GENERATES\ (S_1, S_2, ..., S_n) \rightarrow S_1, S_n,..., S_2$ - $S_2, S_1,..., S_n$ - $S_2, S_n,..., S_1$ -
  $S_n, S_1,..., S_2$ - $S_n, S_2,..., S_1,$ such that $S_i \in N$

It is important to note that $\alpha$ is a chain of characters in $A$ and $\beta$ is a chain of characters in $A^+$. $A^+$ denotes the set of all the words formed from the symbols of the alphabet, excluding the empty word. Now, notice the following example of a generating rule of a *VG*. Given the chain of symbols: $\alpha, \beta, \varphi \in N$ and the generating rule *GENERATES* ($\alpha, \beta, \varphi$), the same produces the following chains of symbols:

$\alpha, \varphi, \beta;\ \ \beta, \alpha, \varphi;\ \ \beta, \varphi, \alpha;\ \ \varphi, \alpha, \beta;\ \varphi, \beta, \alpha$

Rules can also be used in the following manner:

a. Given the chain of symbols: $\alpha$, *GENERATES* ($\beta, \varphi, \omega$) where $\alpha \in A$ and $\beta, \varphi, \omega \in N$

   In this case the symbol $\alpha$ remains fixed to the left and the generating rule will produce combinations of the arguments. The following chains of symbols are obtained: $\alpha, \beta, \omega, \varphi;\ \ \alpha, \varphi, \beta, \omega;\ \ \alpha, \varphi, \omega, \beta;\ \ \alpha, \omega, \beta, \varphi;\ \alpha, \omega, \varphi, \beta$

   Fixed symbols can also be found to the right:
   *GENERATES* ($\beta, \varphi, \omega$), $\alpha$ where $\beta, \varphi, \omega \in N$ and $\alpha \in A$

   Or at both ends: $\alpha$, *GENERATES* ($\beta, \varphi, \omega$), $\sigma$ where $\alpha, \sigma \in A$ and $\beta, \varphi, \omega \in N$

b. Given the chain of symbols:
   *GENERATES* ($\alpha$, *GENERATES* ($\beta, \varphi, \omega$), $\sigma$) where $\alpha, \beta, \varphi, \omega, \sigma \in N$

   The results obtained are as follows: α, β, ω, φ, σ;  α, φ, β, ω, σ;  α, φ, ω, β, σ;  α, ω, φ, β, σ;  α, ω, β, φ, σ; α, σ, β, φ, ω;  α, σ, β, ω, φ; …

The expressive power of the *VGs* is very large. In a single rule we can synthesize several others. The idea is having very general basic rules and generating rules with a very important expressive power. Let's see another example, but now applied to natural languages. Consider this very simple grammar:

*1: S → NS + VS + PS*
*2: NS → NOUM*
*3: VS → VERB*
*4: PS → PREP + NS*

and the sentence: Túpac Amarú luchó en Perú [Tupac Amaru fought in Peru]. The grammar after analyzing the sentence, obtains the following structural description, represented by a syntagmatic tree (Figure 1). If the grammar has a strong generative capability, we can then exchange the syntagms that are at the same structural level and the sentence will continue having syntactic and semantic sense in Spanish. Note that the tree has three syntagms that are at the same level and that can therefore be interchanged. However, the grammar of the example won't be able to recognize them, so would need to add five new rules. With volatile grammars we would keep the basic rules and incorporate one generating rule, if the basic rule cannot be applied. The grammar would be as follows:

*1': S → NS + VS + PS*
*2': S → GENERATES (NS, VS, PS)*
*3': NS → NOUM*
*4': VS → VERB*
*5': PS → PREP + NS*

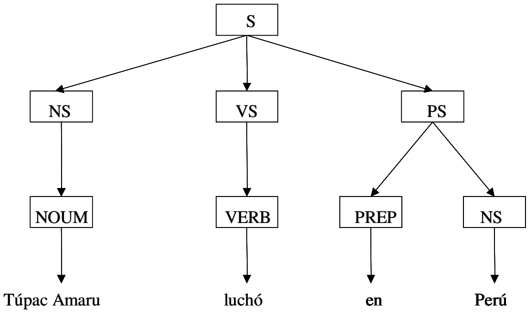


**Fig. 1.** Syntagmatic tree

Finally, it is important to note that *VGs* may easily solve problems of *right recursive,* common in romance languages and English, and of *self-embedding,* which occurs because of the relative pronouns that produce another sentence. Also, *VGs* had to solve some deficiencies that appeared in their *weak generative capacity* and in their *strong generative capacity.* We say that a grammar lacks a weak generative capacity if it cannot generate on its own the grammatically correct sentences. To solve this conflict, a mechanism, formed by a group of rules and restrictions that only allows generating grammatically correct sentences, was developed. On the other hand, we say that a grammar lacks strong generative capacity if it is not able to structurally describe the sentence in a correct manner. This problem is caused, mainly, by ambiguities. To solve this conflict, a series of propositions that solves the ambiguities through punctuation marks was proposed.

## 3 Using Componential Analysis and Integrity Rules in the Interpretation Phase

The semantic theory (ST) of a natural language forms part of the linguistic description of that language. Katz and Fodor [14] state that the linguistic description of a language minus its grammar equals its semantics. We can note that grammar provides identical structural descriptions for sentences with different meanings, and different structural descriptions for sentences with identical meanings. Grammar aspires to describe the structure of a sentence isolated from its context. Semantics, on the other hand, must interpret the sentence within a determined context. The set of rules that constitute grammar is finite, but the set of sentences it can generate is infinite.

Several disciplines, such as philosophy, linguistics, philology, as well as several theorists of semantics, such as Bloomfield, Carnal, Harris, Osgood, Quine, Tarski, Wittgenstein and Ziff, have contributed a vast amount of information about the semantics of natural language. At present, the information is so vast, although in some cases it may be loosely formulated, that the disagreement regarding the creation of a theory of semantics is practically generalized. Chomsky says: part of the difficulty presented by the theory of meaning is that the term *meaning* tends to be used as a catch-all term to include all the aspects of language of what we know very little.

*ST* has two versions, a strong one and a weak one. The strong version states that theory must interpret discourse in the same way in which a fluent speaker would recognize ambiguities, discover anomalous chains, and recognize periphrastic relationships. The weak version is less demanding; it requires the theory to interpret discourse using only grammatical and semantic relationships. In this paper we use the weak version.

The structure of an *ST* can be defined by the following quadruplet: *F = {S, DE, IS, ISC},* where *S* represents the sentence, *DE* the structural description of the sentences, *IS* the semantic interpretation of the sentence and *ISC* the semantic interpretation of the sentence within a particular context. IS and ISC can have multiple representations for the same sentence. If the sentence is *n* times ambiguous in isolation, there will be *n* interpretations in IS. If the sentence, on the other hand, is *n* times ambiguous in a limited context, there will be *n* interpretations in ISC. The sentence may also have *n* interpretations in IS and zero or one in ISC. This would imply that it has no interpretation in a particular context (zero) or that the context eliminated the ambiguity (one).

### 3.1 Componential Analysis

*Componential analysis* ([14], [15]) defines the sense of the lexical parts that constituted the sentence through a set of semantic features. It uses a *data dictionary* and a set of *projection rules.* The data dictionary represents a list of the lexical items that constitute a language, each one associated with a natural access form. Access consists of a grammatical part that provides the classification of the lexical item according to the basic parts of discourse, and a semantic part that provides the different meanings of the lexical item in different contexts. Consider the following French sentence: *Le chat mange la viande [the cat eats the meat].* The dictionary gives the following results for the first two words:

D1-1. *Le* → definite article → [part of the sentence used to denote the extension in which the noun it precedes should be taken]

D2-1. *chat* → noun → (physical object) → (animal) → [meat eating domestic mammal, around 50 cm long from the head to the base of the tail, that measures 20 cm. Round head, rough tongue, short legs, etc.] → <animal>

D2-2. *chat* → noun → (physical object) → (inanimate object) → [purse or bag where money is kept] → <physical object >

D2-3. *chat* → noun → (physical object) → (inanimate object) → (instrument) → [machine composed by a rack and pinion gearing, with a safety ratchet, used to lift heavy weights] → <physical object>

D2-4. *chat* → noun → (dance) → [Argentina; popular dance executed by one or two couples with fast movements] → <social activity>

The *syntactic markers* -noun, verb, adjective, adverb- express the classification of the lexical item according to the basic parts of discourse. The *semantic markers* -physical object, instrument, dance- are described between parentheses. These represent the part of the meaning of the lexical item that is common for the language. The *semantic differentiators* are written between square brackets. These represent the part of the meaning of the lexical item that is not common for the language. Finally, *semantic amalgams* are written between angular brackets. They are used to relate the lexical item with the rest of the vocabulary in the sentence. For each semantic differentiator, there is a semantic amalgam.

It is important to note that in the example provided, the word *chat* [cat] is polysemy, and therefore, from the semantic point of view it has multiple ramifications, which are the cause of semantic ambiguity. Now, a sentence may contain ambiguous lexical items and still not be ambiguous. The word *chat* [cat], for example, has a ramification in the semantic markers: *physical object and dance.* In turn, the marker *physical object* has a new ramification: *animal and inanimate object.* These ramifications are enough to avoid ambiguity in this sentence. In general, an ST may solve ambiguities between semantic markers, but not between semantic differentiators.

The purpose of projection rules is to relate groups of ways dominated by a *grammatical marker*, considering elements that are common to each one, to form a new group of ways that provide a new set of semantic interpretations. Some of the projection rules are presented below.

R1. Given two routes of the form:

P1. Lexical chain 1 → syntactic marker of determiner → semantic markers of determiner → [1]
P2. Lexical chain 2 → syntactic marker of noun → semantic markers of noun → [2]

There is an amalgam of the form:
A1. Lexical chain 1 + Lexical chain 2 → dominant node marker → semantic markers of determiner → [1] → semantic markers of noun → [2] → <2>

R2. Given two routes of the form:

P3. Lexical chain 3 → syntactic marker of verb → semantic markers of verb → [3] → <3>
P4. Lexical chain 4 → syntactic marker of nominal syntagm → rest of the route Lexical chain 4 → <4>

There is an amalgam of the form:
A2. Lexical chain 3 + Lexical chain 4 → dominant node marker → semantic markers of verb → [3] → rest of the route Lexical chain → [4] → <3>

R3. Given two routes of the form:

P5. Lexical chain 5 → syntactic marker of nominal syntagm → rest of the chain Lexical chain 5 → <5>
P6. Lexical chain 6 → syntactic marker of verbal syntagm → rest of the chain Lexical chain 6 → <6>

There is an amalgam of the form:
A3. Lexical chain 5 + Lexical chain 6 → dominant node marker → rest of the route Lexical chain 5 → rest of the route Lexical chain 6

By applying *R1* to *D1* and *D2* we get four derived ways that form *D6.* Note that there are four possible ways and four derived ways. By applying *R1* to *D4* (*la*) and *D5* (*viande*), we get two derived ways that form *D7.* In this case, there are two possible ways and two derived ways. By applying *R2* to *D3* (*mange)* and *D7 (la viande),* we get a derived route that forms *D8.* Note that in this case there are eight possible routes and only one derived route. Finally, by applying *R3* to *D6* (*Le chat*) and *D8* (*mange la viande*), we get one derived route that forms *D9.* Note that in this case there are four possible routes and only one derived route. The route that forms *D9* is the following:

D9-1. *Le chat mange la viande* → sentence → [part of the sentence that serves to denote the extension in which the noun it precedes should be taken] → (physical object) → (animal) → [meat eating domestic mammal, etc.] → (action) → [chewing and finely grinding food etc.] → [part of the sentence that serves to denote the extension in which the noun it precedes should be taken] → (physical object) → (animal) → [soft and fleshy part of the body of animals] → <animal>

The projection rules presented are illustrative of the types of rules used by ST, but they should not be considered a contribution to the STs of romance languages.

### 3.2 Rules of Integrity

*Rules of integrity* have the goal of refining, improving and optimizing translation. These must be specific for each language. There are compression rules, when we translate from Spanish into other romance language and of decompression in the opposite way. The rules are the same, they simply start from the left and are substituted in the right in the first case, and in the second case they start from the right and are substituted from the left. Below are some rules used in French:

RC1: de le → du; RC2 : a le → au; RC3 : de les → des; RC4 : se est → s'est
RC5: la + (word starting with a vowel) → l'(word starting with a vowel)
RC6: la + (word starting with a vowel) → l'(word starting with a vowel)
RC7: verb + la → y + verb
RC8: de + (word starting with a vowel) → d'(word starting with a vowel)

A sentence in Spanish and its translation into French after applying volatile grammars and after being refined by the proofreading rules is presented below. Given the sentence: *Túpac Amarú se pronunció siempre contra el imperio de los españoles [Tupac Amaru always spoke against the empire of the Spaniards]*

The first translation produces:

*Túpac Amarú se est prononcé toujours contre le empire de les espagnols*

and the refinement: *Túpac Amarú s'est prononcé toujours contre l'empire des espagnols*

## 4 Conclusions and Future Work

This paper presents an approach to automatic translation in two steps. The first one is recognition and the last one is interpretation. In the recognition phase, we use volatile grammars. The idea of writing basic utterances and rules that in turn generate syntactic analysis rules, that later disappear saving space, represents a new approach, an innovative and economic form of writing a grammar. The advantages of *GV* are clear: small, easy to parse with, and economic. On the other hand, componential analysis and integrity rules are used mainly in the semantic phase. Componential analysis represents and defines the sense of the lexical parts that constitute the sentence through a collection of semantic features. Integrity rules have the purpose of refining the translation. They work as controllers of the new generated text and must be specific for each language.

The methodology is integrated with up-to-date literature, and seeks to be general, although it is mainly directed to romance languages. We believe that the use by, and feedback from researchers working in this area, will generate ideas, points of views and opinions, which will allow us to discern the strong and weak points of the proposal. The current conceptual framework requires been proved with several romance language to confirm its advantages. It is necessary to translate in both ways from some of them to the others, and vice versa. The results of the exploration will become visible in the near future.

## References

1. Collins, M.: Head-Driven Statistical Models for Natural Language Parsing. Computational Linguistic 29(4), 589–637 (2003)
2. Klein, D., Manning, C.: Accurate Unlexicalized Parsing. In: Proc. of the 41st Annual Meeting of the Association for Computational Linguistics, July 2003, pp. 423–430 (2003)
3. Chang, P., Toutanova, K.: A Discriminative Syntactic Word Order Model for Machine Translation. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pp. 9–16 (2007)
4. Ablanedo, J., Aiken, M., Vanjani, M.: Efficacy of English to Spanish automatic translation. International Journal of Information and Operations Management Education 2(2), 194–210 (2007)
5. Nirenburg, S., Somers, H., Wilks, Y.: Reading in Machine Translation. MIT Press, Cambridge (2003)
6. Menezes, A., Quirk, C.: Using Dependency Order Templates to Improve Generality in Translation. In: Proceedings of the Second Workshop on Statical Machine Translation. Association for Computational Linguistics, pp. 1–8 (2007)
7. Colmerauer, A.: Metamorphosis Grammars. Lectures Notes in Computer Science, vol. 63, pp. 133–189. Springer, Heidelberg (1978)
8. Pereira, F., Warren, D.: Definite Clause Grammar for Language Analysis. –A survey of the formalism and a comparison with augmented transition network. Artificial Intelligence 13, 231–278 (1980)
9. Pereira, F.: Extraposition Grammars. Computational Linguistic 7(4), 243–256 (1981)
10. Dahl, V.: Discontinuous Grammars. Computational Intelligence 5(5), 161–179 (1989)
11. Strzalkowski, T.: Reversible Grammar in Natural Language Processing. Kluwer Academic Publishers, Dordrecht (1993)
12. Cabasino, T., Paolucci, P., Todesco, G.: Dynamic Parsers and Evolving Grammars. ACM SIGPLAN Notices 27(11), 39–48 (1992)
13. Searls, D.: String variable grammar: a logic grammar formalism for the biological language of DNA. The Journal of Logic Programming 24(1), 73–102 (1995)
14. Katz, J., Fodor, J.: The Structure of a Semantic Theory. Language 39, 170–210 (1963)
15. Nirenburg, S., Raskin, V.: Ontological Semantics. MIT Press, Cambridge (2004)

# Representation of Uncertain Knowledge in Probabilistic OLAP Model

Maciej Kiewra

Institute of Information Science and Engineering, Wroclaw University of Technology,
Janiszewskiego 11/17, 50-370 Wrocław, Poland
maciej.kiewra@pwr.wroc.pl

**Abstract.** The probabilistic OLAP model has been presented in this paper. It permits uncertain knowledge to be represented in data warehouse systems. There are two types of uncertainty that can be expressed in this model: imprecise facts and uncertain facts. The former are facts that have occurred but their characteristics are not certain. The latter are facts whose occurrences are uncertain. Typical OLAP algebra operators (set operators, restriction, projection etc.) are included in this model.

**Keywords:** data warehouse, OLAP, knowledge management, uncertainty, probabilistic databases, multidimensional model.

## 1 Introduction

Data warehouse technology and in particular OLAP model (OLAP stands for On-Line Analytical Processing) is an industry accepted standard for storage and management of knowledge that is crucial for decision making processes. Although this type of knowledge very often possesses uncertain character, uncertainty is typically eliminated during data cleansing and transformation. For instance, consider a multidimensional database related to international money laundering in which one of the attributes is the origin country. Let's assume that it is not possible to determine the exact origin country for many facts from this database, because the available information indicates only that the illegal transactions probably comes from Chile or Peru. In the traditional approach, an origin country dimension would have the hierarchical structure (Continent - Country) where for each continent the "unknown" element would be included. Obviously, the information that probable origin country is Chile or Peru would be replaced by *unknown* and irreversibly lost in the traditional approach.

The example presented above describes the fact that will be denominated an *imprecise fact* in this paper (We are sure that the fact has occurred but its characteristic is uncertain). Another type of uncertainty is related to the occurrence of the fact. Let's imagine that the suspect transaction has been detected and as a result of investigations two hypothesis have been elaborated. The first one, slightly more probable, assumes that the encountered operations are related to illegal business activities , the second one indicates that the transaction

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 89–96, 2008.

is legal (in this case the fact should not be included into the money laundering database). In the traditional approach, the arbitrary decision must be taken (the fact should be included or not) and uncertainty is eliminated. This type of facts will be denominated *uncertain facts* in this paper (it is not sure if the fact of money laundering has occurred at all).

It is easy to notice that uncertain knowledge (knowledge composed of uncertain and imprecise facts) cannot be stored and managed in traditional data warehouse systems without the risk of serious inconsistency. The purpose of this paper is to present a multidimensional algebra that will be able to represent two types of uncertainty mentioned above by means of probability.

## 2 Related Works

The acronym OLAP was introduced by Codd in [2]. Among all attempts of formalizing of OLAP model that appeared in late 90's, the most representative is [3]. Although one of the first works that introduced uncertainty in the OLAP model were published in the same period [4], there are few works concerning uncertain knowledge in OLAP model. In [4] it assumed that the facts in fact tables can be related to dimensions on different granularity levels (the lowest level represented precise data, the rest of them imprecise ones). The model from [4] can store only imprecise facts (e.g. $5,000,000 has been transfered from South America - the origin country is unknown). Additionally, [4] does not store estimated probability directly in the model and it is focused on aggregation of imprecise data without defining other operators typical in the OLAP algebra [3].

The uncertain model from [1] presents the widest set of operators (restriction, projection, among others). Uncertain facts are included in this model and they are expressed by means of probability. The probability of the fact occurrence is defined at the row level so there is no way to represent imprecise facts. Moreover, in [1] it is assumed that two rows from the fact table that differ only in the probability value must concern the same fact. Consider the cube from the Example 1 presented below. The model from [1] is not able to store two facts that will be reflected the separate attempts related to transfer of $5,000,000 from Colombia in February 2007. Moreover, dimension hierarchies are not explicitly defined in this paper.

In [5] imprecise facts are represented (this concept is very similar to [4]) and probability distributions are used to represent the measure values. Although this distributions reflect uncertainty, the uncertainty corresponds to measure values – the occurrence of facts is certain. In other words, it is not possible to represent uncertain facts in this model. Similarly to [4], OLAP algebra operators are not defined in [5]. Moreover, conditions introduced by the authors would make difficult to introduce the restriction operator and set operators in general.

Enumerating all attempts to modeling uncertain knowledge in OLAP by means of probability, [7] should be mentioned (although it does not define OLAP algebra and it is tightly connected with a real-world case study of location-based services - LBS).

It is important to emphasize that probabilistic OLAP models benefit from researches on probabilistic databases. In [8] relational algebra was extended by the probability theory in order to express uncertainty about object properties.

Finally, fuzzy logic also has been used to handle uncertainty in OLAP model [6] (fuzzy hierarchies were introduced in dimension structures).

## 3 Probabilistic Model

**Definition 1.** *Let $d = (id_d, A_d, E_d)$ be a dimension, where:*

- *id is a dimension identifier*
- *$A_d = (a_0, a_1, \ldots, a_n)$ is a sequence of dimension attributes that constitute a hierarchy $a_0$ is the root attribute and $a_i$ is an attribute of the dimension $d$ on the $i^{th}$ level (for $i = 1, \ldots, n$).*
- *$E_d = (E_0, E_1, \ldots, E_n)$ is a sequence of sets of attribute members (elements) $E_0 = \{All, \emptyset\}$ and $E_i = \{(e, e_p) : e \in dom(a_i) \wedge e_p \in dom(a_{i-1})\}$ where: $dom(a_i)$ is the domain of the attribute $a_i$ (the set of all possible values of $a_i$) for $i = 1, \ldots, n$; $e$ is the unique name of a given member and $e_p$ is the name of its parent from the upper level. Additionally, let $dom_l(d)$ be the domain of the $n^{th}$ attribute (the leaf attribute) and $dom(d) = \bigcup_{i=1}^{n} dom(a_i)$*

As it can be observed, Definition 1 groups all attributes from a single dimension into a hierarchy. If it is impossible to form a hierarchy for two attributes of a single dimension, this dimension should be splitted into two separate dimensions.

The restriction operator defined below uses the *anc* function:

Let anc: $dom(a_i) \rightarrow dom(a_j)$ for $(n \geq i > 0, i > j \geq 0)$ be a function that for every member of the attribute ai returns its ancestor at the level of $a_j$. The function value can be calculated recursively from the following formula:

$$(\underset{i=j+1,(e_i,e_p)\in E_i}{\forall} anc(e_i) = e_p) \wedge (\underset{i>j+1,(e_i,e_p)\in E_j}{\forall} anc(e_i) = anc(e_p)) \qquad (1)$$

**Definition 2.** *Let $C = (D, M, V)$ be a cube where: $D = \{d_1, \ldots, d_N\}$ is a set of dimensions used in the cube $C$, $M = \{(m_1, f_1), \ldots, (m_{Nm}, f_{Nm})\}$ is the set of numeric measures stored in the cube $C$, where $m_i$ is a unique measure name and $f_i$ (for $i = 1, \ldots, N_m$) is an aggregate function that will be used to obtain aggregated values, $V = \{v_1, \ldots, v_{Nv}\}$ is a set of cube cells (facts) - each cell is a tuple: $v_i = (id_i, p_i, V^d_{i1}, \ldots, V^d_{iN}, V^m_i)$ for $i = 1, \ldots, N_v$ where:*

- *$id_i$ is the fact identifier. If $id_i = 0$ then it is aggregated fact (see below),*
- *$p_i$ is the probability that the fact $v_i$ has occured.*
- *$V^d_{ij} = \{(v^d_{ij}, p^d_{ij}) : v^d_{ij} \in dom_l(d_j) \wedge \sum_{k=1}^{card(V^d_{ij})} p^d_{ijk} <= 1\}$ is a subset of a set that represents the discrete distribution of a random variable related to the value of dimension $d_j$ (for $j = 1, \ldots, N$) in the fact $v_i$; $v^d_{ijk} \in dom_l(d_j)$ is*

*one of possible values;* $p^d_{ijk}$ *is the conditional probability that the fact* $v_i$ *has associated the value* $v^d_{ijk}$ *given the fact represented by* $v_i$ *has occurred. For simplicity, if* $V^d_{ij} = \{v^d_{ij1}, 1)\}$*, then it will be denoted* $V^d_{ij} = v^d_{ij}$

- $V^m_i = (v^m_{i1}, \ldots, v^m_{iNm})$ *is a sequence of measure values, where* $v^m_{ij}$ *is a value of the measure* $m_j$ *in the fact* $v_i$ *(for* $j = 1, \ldots, N_m$*);*

The Definition 2 deserves a few comments. First of all, introduction of the identifier $id_i$ facilitates set operations between cubes defined below. Since the same real world fact expressed in two cubes can possess different values for the same dimensions, the mechanism to identify the same facts is indispensable. As it can be seen, there are two symbols related to probability that reflects two types of uncertainty mentioned in the first section: $p^d_{ijk}$ - permits imprecision to be reflected while $p_i$ is used to represent uncertain facts.

If $\underset{i,j}{\exists} \sum_{k=1}^{card(V^d_{ij})} p^d_{ijk} < 1$, it means that distribution of the random variable related to the dimension $d_j$ in the fact $v_i$ is known partially. This phenomena is denominated missing probability (e.g. [9]).

**Example 1.** Consider the cube related to money laundering $C_1 = (D_1, M_1, V_1)$ where $D_1 = \{d_1, d_2, d_3\}$ $d_1 =(ORIGIN, (total, ocontinent, ocountry), (\{(All, \emptyset)\},$ $\{(NorthAmerica, All)$,(South America, All),...},
{(Antigua and Barbuda,North America),...}));
$d_2 = (DESTINATION, (total, dcontinent, dcountry), (\{(All, \emptyset)\},$
{(North America,All),(South America, All),...},{(Antigua and Barbuda,North America),...}));
$d_3 = (PERIOD, (total, year, month), (\{(All, \emptyset)\}, \{(2007, All)\}, \{(jan, 2007), \ldots,$ $(dec, 2007\})); M_1 = \{(amount, sum), (operation - number, count)\}$ and the set $V_1$ is given by the Table 1.

**Table 1.** The set $V_1$

| $id_i$ | $p_i$ | PERIOD | ORIGIN | DEST | AMOUNT | OPER NUM. |
|---|---|---|---|---|---|---|
| 1 | 1.0 | Jan 2007 | Colombia | France | $5,000,000 | 1 |
| 2 | 0.5 | Jan 2007 | Mexico | Italy | $2,000,000 | 1 |
| 3 | 1.0 | Feb 2007 | {(Peru,0.5),(Chile,0.25),(Cuba,0.25)} | USA | $1,000,000 | 1 |
| 4 | 0.9 | Feb 2007 | {(Colombia,0.8),(Peru,0.1)} | Spain | $4,000,000 | 1 |

There is no uncertainty in the first fact. The second fact is uncertain (the probability that $2,000,000 has been transferred illegally from Mexico is equal 0.5). The third fact is certain, but imprecise (the exact origin is unknown, but Peru is the most probable option). The last fact contains two types of uncertainty and the distribution of the random variable related to the ORIGIN dimension is known partially.

### 3.1 Set Operators

Before set operators can be defined, equivalence of facts will be introduced.

**Value-equivalence.** The two facts represented by $v_1 = (id_1, p_1, V_{11}^d, \ldots, V_{1N}^d, V_1^m)$ and $v_2 = (id_2, p_2, V_{21}^d, \ldots, V_{2N}^d, V_2^m)$ are value equivalent $(v_1 \equiv v_2)$ if the following two conditions are fulfilled:

$$id_1 = id_2 \wedge p_1 = p_2 \wedge V_1^m = V_2^m \wedge \underset{N \geq i > 0}{\forall} (\underset{e \in dom(d_i)}{\exists} ((e, p_1^d) \in V_{i1}^d \wedge (e, p_2^d) \in V_{i2}^d)) \Rightarrow p_1^d = p_2^d) \tag{2}$$

$$v_1 \equiv v_2 \Rightarrow \underset{N \geq i > 1}{\forall} (\sum_{k=1}^{card(V_{1i}^d \cup V_{2i}^d)} p_k^d \leq 1) \tag{3}$$

According to (2) two value-equivalent facts can differ only in the elements of sets: $V_{1i}^d$ and $V_{2i}^d$ (for $N \geq i > 0$). Additionally, common elements from this two sets must possess equal probability values. The formula (3) guarantees that the fact, that will be created as a result of union of two value equivalent facts, will hold the inequality from the Definition 2. Only value-equivalent tuples are coalesced in union and intersection operations. If tuples are not value-equivalent, they are treated as completely different facts.

**Intersection.** let $C_1 \cap C_2 = C_0$ be the intersection operator between the cubes $C_1 = (D_1, M_1, V_1)$ and $C_2 = (D_2, M_2, V_2)$. The result cube $C_0 = (D_0, M_0, V_0)$ holds: $D_0 = D_1 = D_2$, $M_0 = M_1 = M_2$. $V_0 = V_1 \cap V_2$. is calculated using the following formula:

$$\underset{v_0 \in V_0}{\forall} \underset{v_1 \in V_1, v_2 \in V_2}{\exists} v_1 \equiv v_2 \wedge id_0 = id_1 \wedge p_0 = p_1 \wedge V_0^m = V_1^m \wedge \underset{N \geq i > 0}{\forall} V_{0i}^d = V_{1i}^d \cap V_{2i}^d) \tag{4}$$

**Union.** let $C_1 \cup C_2 = C_0$ be the union operator between the cubes $C_1 = (D_1, M_1, V_1)$ and $C_2 = (D_2, M_2, V_2)$. The result cube $C_0 = (D_0, M_0, V_0)$ holds: $D_0 = D_1 = D_2$, $M_0 = M_1 = M_2$ and $V_0 = V_1 \cup V_2$. Value equivalent facts are coalesced according to the following condition:

$$\underset{v_1 \in V_1, v_2 \in V_2}{\exists} \Rightarrow \underset{v_0 \in V_0}{\exists} id_0 = id_1 \wedge p_0 = p_1 \wedge V_0^m = V_1^m \wedge \underset{N \geq i > 0}{\forall} V_{0i}^d = V_{1i}^d \cup V_{2i}^d \tag{5}$$

**Difference.** let $C_1 - C_2 = C_0$ be the difference operator between the cubes $C_1 = (D_1, M_1, V_1)$ and $C_2 = (D_2, M_2, V_2)$. The result cube $C_0 = (D_0, M_0, V_0)$ holds: $D_0 = D_1 = D_2$, $M_0 = M_1 = M_2$ and $V_0 = V_1 - V_2$ where each $v_0 \in V_0$ holds one of the following formulas:

$$1. v_0 \in V_1 \wedge (\neg \underset{v_2 \in V_2}{\exists} v_2 \equiv v_0) \tag{6}$$

$$2. \underset{v_1 \in V_1, v_2 \in V_2 \wedge v_2 \equiv v_1}{\exists} \underset{N \geq i > 0}{\exists} card(V_{1i}^d - V_{2i}^d) > 0) \wedge v_0 = d(v_1 - v_2) \tag{7}$$

The difference between value-equivalent tuples is obtained using $d$ function:

$$\underset{v_1 \equiv v_2}{\forall} d(v_1, v_2) = v_0 \wedge \underset{N \geq i > 0 \wedge V_{1i}^d - V_{2i}^d \neq \emptyset}{\forall} (V_{0i}^d = V_{1i}^d - V_{2i}^d) \wedge \underset{N \geq i > 0 \wedge V_{1i}^d - V_{2i}^d = \emptyset}{\forall} (V_{0i}^d = V_{1i}^d) \tag{8}$$

### 3.2 Restriction, Projection and Aggregation

The last element needed for the definition of the restriction operator is the set of valid expressions that can be used in order to perform the restriction.

**Expression Set.** $R_C$ is a set of expressions that restricts cells of the cube $C$. The set $R_C$ is built according to the following rules:

1. $(a = s) \in R_c$ where $\underset{d \in D}{\exists} a \in A_d$ and $s$ is an unrestricted string
2. $(p \in p_v) \in R_c$ where $p_v \in (0, 1]$
3. $\underset{r \in R_C}{\forall} \neg r \in R_C$
4. $\underset{r_1, r_2 \in R_C}{\forall} (r_1 \wedge r_2) \in R_C \wedge (r_1 \vee r_2) \in R_C$

**Restriction.** Let $\sigma_r(C_1) = C_0$ be the restriction operator that restricts the cube $C_1 = (D_1, M_1, V_1)$ according to the expression $r \in R_{C1}$ and returns the cube $C_0 =< D_0, M_0, V_0 >$ where: $D_0 = D_1$, $M_0 = M_1$. $V_0$ is built according to the following rules:

1. $\underset{r=(a=s)}{\forall} (\underset{N>=i>0}{\exists} a = a_{card(A_{d_{0i}})} \wedge \underset{v_1 \in V_1}{\exists} v^d_{1i} = s) \Rightarrow v_1 \in V_0)$
2. $(\underset{r=(a=s)}{\forall} \underset{N \geq i>0}{\exists} \underset{card(A_{di})>j>0}{\exists} a = a_j \wedge \underset{v_1 \in V_1}{\exists} v^d_{1i} = anc(s)) \Rightarrow v_1 \in V_0$
3. $\underset{r=(p>=p^v)}{\forall} \underset{v_1 \in V_1}{\forall} (p_1 \geq p^v \wedge \underset{card(N) \geq j>0 \wedge card(V^d_{1j}) \geq k>0}{\forall} p^d_{1jk} \geq p^v \Rightarrow v_1 \in V_0)$
4. $\underset{r \in R_{C_1}}{\forall} \neg\sigma_r(C_1) = C_1 - \sigma_r(C_1)$
5. $\underset{r, r_1 \in R_{C_1}}{\forall} \sigma_{r \wedge r_1}(C_1) = \sigma_r(C_1) \cap \sigma_{r_1}(C_1)$
6. $\underset{r, r_1 \in R_{C_1}}{\forall} \sigma_{r \vee r_1}(C_1) = \sigma_r(C_1) \cup \sigma_{r_1}(C_1)$

The rules used to built the set $V_0$ need a few comments. The first rule restricts the cube to the cells that possess the value $e$ in the leaf attribute of the dimension $d_i$ (e.g. *country=Peru*). The second one restricts the cube to the cells that have the value $e$ in non-leaf attributes of the dimension $d_i$ (e.g. *continent=Europe*). The third one restricts the cells whose probability of occurrence is greater or equal $p^v$. The rest of rules is used to define the restriction for negation, conjunction and alternative of expressions.

**Projection.** Let $\Pi_{M'}(C_1) = C_0$ be the projection operator that limits the set of measures from the cube $C_1 = (D_1, M_1, V_1)$ to the subset $M' \subseteq M_1$ and returns the cube $C_0 =< D_0, M_0, V_0 >$ where: $D_0 = D_1$, $M_0 = M'$ and $V_0 = \{v_0 : v_0 = (id_0, p_0, V^d_{01}, ..., V^d_{0N}, (v^m_{01}, \ldots, card(v^m_{0card(M')})))\}$ holds the following condition:

$$\underset{v_0 \in V_0}{\forall} \underset{v_1 \in V_1}{\exists} id_0 = id_1 \wedge p_0 = p_1 \wedge \underset{N \geq i>0}{\forall} V^d_{0i} = V^d_{1i} \wedge \underset{m \in M'}{\forall} v^m_0 = v^m_1 \qquad (9)$$

Projection is quite simple, intuitive and very similar to its relational counterpart. It limits the set of measure to the given subset.

**Aggregation.** Let $\alpha_G(C_1) = C_0$ be the aggregation operator that aggregates the tuples from the cube $C_1 = (D_1, M_1, V_1)$ using attributes from the set $G$ and returns the cube $C_0 = (D_0, M_0, V_0)$ where: $M_0 = M_1$. The set $D_0$ is built using the following condition:

$$\underset{d_0 \in D_0}{\forall} \Big( \underset{d_1 \in D_1 g \in G}{\exists} g \in A_{d_1} \wedge id_{d_0} = id_{d_1} \wedge A_{d_0} = G \cap A_{d_1} \wedge dom(d_0) = dom(d_1) \cap dom(G) \Big) \tag{10}$$

and the set $V_0 = \{v_0 : v_0 = (0, 1.0, v_1^d, \ldots, v_{card(D_0)}^d, E_{m1}, \ldots, E_{m_{card(M_0)}})\}$ where: $v_i^d \in dom_l(d_i)$ is the value of the dimension $d_i \in D_0$ in the aggregated fact $v_0$ for $i = 1, \ldots, card(D_0)$; $E_j^m$ is the expected value of the random variable related to the aggregated value of the measure $m_j$ in the fact $v_0$, for $j = 1, \ldots, card(M_0)$. The aggregation operation has the following intuition - aggregated data are the data that should be presented to the user. Since millions of cells can be involved in a single aggregation, maintaining aggregated probability distribution may give enormous collection of data - very difficult to analyze and understand. For this reason, expected values are calculated for aggregated value of each measure and involved probabilities are omitted.

Due to space limitations less important and obvious in definition OLAP algebra operators have been omitted (e.g. rename, force and extract [1,3])

**Example 2.** Consider the cube from the Example 1 and two following queries:

1. Non aggregated data related to money transfers that does not come from Colombia
2. The total amount of money transferred to Europe in Jan. 2007 grouped by origin countries omitting Peru

The Table 2 contains sample queries with the results.

**Table 2.** The sets from the Example 2

| $\sigma_{\neg country=Colombia}(C_1)$ | | | | | | |
|---|---|---|---|---|---|---|
| $id_i$ | $p_i$ | PERIOD | ORIGIN | DEST | AMOUNT | OPER NUM. |
| 2 | 0.5 | Jan 2007 | Mexico | Italy | $2,000,000 | 1 |
| 3 | 1.0 | Feb 2007 | {(Peru,0.5),(Chile,0.25),(Cuba,0.25)} | USA | $1,000,000 | 1 |
| 4 | 0.9 | Feb 2007 | {(Colombia,0.8),(Peru,0.1)} | Spain | $4,000,000 | 1 |

| $\alpha_{\{ocountry\}}(\sigma_{continent=''Europe''\wedge\neg ocountry=Peru}(\Pi_{\{amount\}}(C_1)))$ | | | |
|---|---|---|---|
| $id_i$ | $p_i$ | ORIGIN | AMOUNT |
| 0 | 1.0 | Colombia | $7,880,000 |
| 0 | 1.0 | Mexico | $1,000,000 |

## 4 Conclusion and Future Work

The probabilistic OLAP model has been presented in this paper. It represents uncertain knowledge in data warehouse systems. The most important contribution of the paper is the presented model, that to the best of belief of the author is:

- The first one that defines typical OLAP algebra operators (such as restriction, projection, aggregation etc.) for imprecise facts.
- The first one that combines uncertain facts (the facts whose occurrence is not certain) with imprecise facts.
- The first one that defines OLAP algebra operators for uncertain hierarchical data.

The future work will be concentrated on elaborating an effective method for calculating expected values in the aggregation operator. Additionally, the methodology of estimating fact probabilities using data stored in cubes will be developed. Another interesting challenge is related to extension of the model with uncertain measure values and introduction of uncertainty in dimension hierarchies.

## References

1. Moole, B.R.: A Probabilistic Multidimensional Data Model and Algebra for OLAP in Decision Support Systems. In: IEEE Southeast Con., pp. 18–30. IEEE Computer Society Press, Los Alamitos (2003)
2. Codd, E.F.: Providing OLAP to user-analysts: An IT mandate. E.F. Codd and Associates (1993)
3. Datta, A., Thomas, H.: The cube data model: a conceptual model and algebra for on-line analytical processing in data warehouses. Decision Support Systems 27(3), 289–301 (1999)
4. Pedersen, T.B., Jensen, C.S., Dyreson, C.E.: Supporting Imprecision in Multidimensional Databases Using Granularities. In: 11th International Conference on Scientific on Scientific and Statistical Database Management, pp. 90–101. IEEE Computer Society, Los Alamitos (1999)
5. Burdick, D., Deshpande, P.M., Jayram, T.S., Ramakrishnan, R., Vaithyanathan, S.: OLAP over uncertain and imprecise data. The VLDB Journal 16(1), 123–144 (2007)
6. Delgado, M., Molina, C., Sanchez, D., Vila, A., Rodriguez-Ariza, L.: Fuzzy multidimensional model for supporting imprecision in OLAP. In: Fuzzy Systems, pp. 1331–1336. IEEE Computer Society, Los Alamitos (2004)
7. Timko, I., Dyreson, C.E., Pedersen, T.B.: Probability Distributions as Pre-Aggregated Data in Data Warehouses. Technical Report (2005), `http://www.cs.aau.dk/DBTR`
8. Dey, D., Sarkar, S.: A probabilistic relational model and algebra. ACM Trans. Database Systems 4(4), 397–434 (1996)
9. Pittarelli, M.: An Algebra for Probabilistic Databases. IEEE Transactions on Knowledge and Data Engineering 6(2), 293–303 (1994)

# Local Topology of Social Network Based on Motif Analysis

Krzysztof Juszczyszyn, Przemysław Kazienko, and Katarzyna Musiał

Wroclaw University of Technology,
Wyb. Wyspianskiego 27, 50-370 Wroclaw, Poland
{krzysztof, kazienko, katarzyna.musial}@pwr.wroc.pl

**Abstract.** Network motifs – small subgraphs that reflect local topology can be used to discover general profile and properties of the network. Analysis of motifs for the large social networks derived from email communication is presented in the paper. The distribution of motifs in all analyzed real social networks is very similar one another and can be treated as the network fingerprint. This property is most distinctive for stronger human relationships.

## 1 Introduction

When investigating the topological properties and structure of complex networks we must face a number of complexity–related problems. In large social networks, tasks like evaluating the centrality measurements, finding cliques, etc. require significant computing overhead. In this context the methods, which proved to be useful for medium and small networks fail when applied to very large structures. This also refers to complex biological or technology–based networks like computer networks, WWW, gene transcription networks.

The outcomes of the research on network motif detection and analysis in large email–based social network of the Wroclaw University of Technology, consisting of over 5,700 nodes and 140,000 edges (Fig. 1) are presented in this paper. The local structure of this large social network has been investigated by analyzing interconnection patterns between small sets of nodes, called motifs. These small motifs reflect general local topology profile and the properties of the entire network.

## 2 Related Work

Complex networks, both biological and engineered, were analyzed with respect to so–called *network motifs* [6]. They are small (usually 3 to 7 nodes in size) subgraphs which occur in the given network far more (or less) often then in the equivalent random networks (in terms of the number of nodes, node degree distribution, average path length, clustering, etc). Despite all these structural and statistical similarities, networks from different fields have very different local topological structure. It was recently shown that concentration of network motifs may help to distinguish and

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 97–105, 2008.

classify complex biological, technical and social networks [9]. We can distinguish so–called *superfamilies* of networks [9], which correspond to the specific *significance profiles* (SPs). To create SP for the given network, the concentration of individual motifs is measured and compared to their concentration in a number of random networks. The statistical significance of motif *M* is defined by its Z–score $Z_M$:

$$Z_M = \frac{n_M - \left\langle n_M^{rand} \right\rangle}{\sigma_M^{rand}} \tag{1}$$

where $n_M$ is the frequency of motif *M* in the given network, $\left\langle n_M^{rand} \right\rangle$ and $\sigma_M^{rand}$ are the mean and standard deviation of *M*'s occurrences in the set of random networks, respectively [3]. Most algorithms for detecting network motifs assume exhaust enumeration of all subgraphs with a given number of nodes in the network. Their computational cost dramatically increases with the network size. However, it was recently show that it is possible to use random sampling to effectively estimate concentrations of network motifs. The algorithm presented in [4] is asymptotically independent of the network size and enables fast detection of motifs in very large networks with hundreds of thousands of nodes and larger.

The existence of network motifs affects not only topological but also functional properties of the network. For biological networks, it was suggested that network motifs play key information processing roles [11]. For example, so–called FFL motif – Feed–Forward Loop (motif no. 5 in Fig. 2) has been shown both theoretically and experimentally to perform tasks like sign–sensitive filtering, response acceleration and pulse–generation [7]. Such results reveal that, in general, we may conclude about function and properties of very large networks from their basic building blocks [8].

In another work, motif analysis was proved to have ability of fast detection of the small–world and clustering properties of the large network [2]. This result open promising but still unexplored possibilities of reasoning about network's global properties with sampling of local topological structures.

Very little research has been done on motifs in computer science and sociology. SPs for small social networks (<100 nodes) were studied in [9]. A web network counting 3.5x10$^5$ nodes [1] was used to show the usability of sampling algorithm [4].

## 3 Motif Analysis Applied to Large Social Network

To discover properties of large social networks using the motif analysis, some important issues should be respected. The key parameter that reflects the significance of motifs is Z–score (Eq. 1). It is based on comparing the actual concentration of subgraphs (motifs) in the considered network with their concentration in a set of random networks. The size of this set should be as small as possible, so we determined what number of random networks is required to detect motifs with the given accuracy. The actual profile of the network is expressed by the set of Z–scores of the motifs. In our case, we checked all the directed three–node subgraphs (triads). Their concentration values for all triads form so–called *Triad Significance Profile* of the network (TSP) [9]. An email-based social network is the directed, weighted graph so it differs from

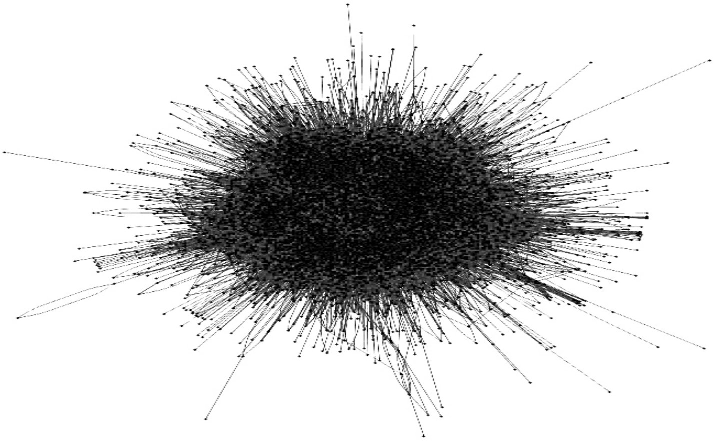


**Fig. 1.** Social network discovered from the email communication between employees of WUT

WWW, gene transcription or molecular networks (unweighted graphs). The weights of the edges in the social network depend on the intensity of communication. However, to enable analysis the domain of edge weights need to be discretized – only small set of classes can be analyzed.

### 3.1 Extraction of the Social Network from Email Communication

The experiments were carried out on the logs from the Wrocław University of Technology (WUT) mail server, which contain only the emails incoming to the staff members as well as organizational units registered at the university (Fig.1) [5]. All experiments were performed with FANMOD tool [13, 14] dedicated for motif detection in large networks.

First, the data cleansing process was executed. The bad email addresses was removed from the analysis and the duplicated ones were unified. Additionally, only emails from and to the WUT domain were left.

Note that although every single email provides information about the sender activity, it can simultaneously be sent to many recipients. An email sent to only one person reflects strong attention of the sender directed to this recipient. The same email sent to 10 people does not respect each individual recipient so much. For that reason, the strength of email communication $S(x, y)$ from email user $x$ to $y$ has been defined in the following way:

$$S(x,y) = \sum_{i=1}^{card(EM(x,y))} \frac{1}{n_i(x,y)}, \tag{2}$$

where: $EM(x,y)$ – the set of all email messages sent by $x$ to $y$; $n_i(x,y)$ – the number of all recipients of the $i$th email sent from $x$ to $y$. In consequence, every email with more than one recipient is treated as $1/n$ of a regular one ($n$ is the number of its recipients).

The strength of the relationship *RS*(*x*,*y*) between *x* and *y* is calculated as follows:

$$RS(x,y) = \frac{S(x,y)}{n(x)}, \tag{3}$$

where: *n*(*x*) – the total number of emails sent by user *x*, was introduced. The values of this function are from range [0,1]. The similar approach was utilized by Valverde et al. to calculate the strength of relationships. It is established as the number of emails sent by one person to another [12]. However, the authors do not respect the general activity of the given individual. In our approach, this general, local activity exists in the form of denominator in Eq. 3. Based on the *RS* values the email-based social network has been created. The next step of preprocessing is to convert continuous weights of ties, i.e. *RS*(*x*,*y*) into five classes (Table 1). The ranges of *RS* (classes) were established to balance the number of edges in each class.. Note also that every node in the network (except only one node in class 2 and 5) belongs to every class since it is incident to at least one edge from every single class. Having classes extracted, the separate networks were built based on the edges from one or more classes. These networks were used for motifs detection in order to check how the TSP profile changes when different communication intensity is considered.

The general statistics about the extracted classes are presented in Table 1. Note that in the class 1 where the strength of the relations is the lowest, the contribution of mutual edges is the smallest – only 1.2% whereas this rate for class 5 – only the strongest connections (*RS*>0.05) was as much as 16.2%. It means that stronger human relationships tend to be more frequently mutual compared to the weaker ones.

Each of the classes separately as well as their various combinations were utilized in the process of detecting triads within the WUT email social network. There are 13 different motifs that consist of three nodes each (Fig. 2). Their ID=1,2,…,13 are used in the further descriptions interchangeably with the corresponding M1, M2,…, M13.

**Table 1.** The number of nodes and edges in the particular classes

| | Class 1 | Class 2 | Class 3 | Class 4 | Class 5 | All |
|---|---|---|---|---|---|---|
| Range of *RS* | (0;0.001] | (0.001;0.005] | (0.005;0.01] | (0.01;0.05] | (0.05;1] | (0;1] |
| No. of nodes | 5,783 | 5,782 | 5,783 | 5,783 | 5,782 | 5,783 |
| No. of single edges (% of total in class) | 30,788 (98.8%) | 33,755 (91.6%) | 18,800 (93.6%) | 28,249 (85.2%) | 13,039 (83.8%) | 124,631 (91.1%) |
| No. of mutual edges (% of total in class) | 382 (1.2%) | 3,086 (8.4%) | 1,283 (6.4%) | 4,919 (14.8%) | 2,529 (16.2%) | 12,199 (8.9%) |
| Total no. of edges | 31,170 | 36,841 | 20,083 | 33,168 | 15,568 | 136,830 |

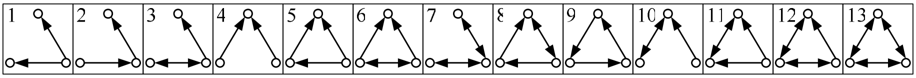


**Fig. 2.** Directed triads and their IDs that can exist within the social network

### 3.2 Triad Significance Profile in Relation to Number of Random Networks

The goal of the first phase of the experiments was to determine the minimum number of random networks which allow to detect the motifs with the required accuracy.

Triad Significance Profiles (TSPs) for all motifs (Eq. 1) were computed separately for the different numbers of random networks (RN), Fig. 3. Note that there are only little differences between obtained Z–scores for the numbers of random networks above 50. It appears that 100 random networks provide sufficient accuracy of calculations. To be sure 500 random networks were used in further research.

We may also conclude, that the considered network reveals the typical property of social networks – the small–world phenomenon. Loosely connected motifs with only 2 edges, like M2, M3, M4, M7, M10 occur less frequently compared to the random networks. It obviously proves the high clustering level, i.e. high probability that two neighboring nodes have connected neighbors. The only exception is M1 which is met relatively often. This reflects specific property of large mail–based social network: there are relatively many broadcasting nodes who spread messages (news, announcements, bulletins) which are never answered.

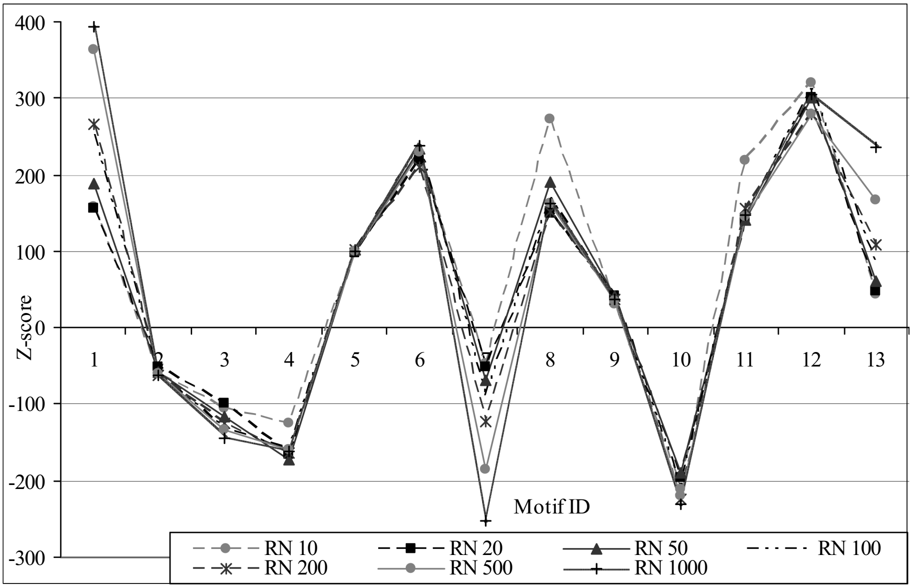


**Fig. 3.** Triad Significance Profile (Z-score values) of the WUT email-based social network for different numbers of random networks

### 3.3 Triad Significance Profile in Relation to Strength of the Ties

Triad Significance Profile for separated and aggregated classes of ties' strengths are presented in Table 2 and Fig. 4. "Class 12345" stands for entire network (all classes merged), while "Class 5" denotes only the subnetwork created by the strongest ties

**Table 2.** Z–score values for particular motifs in the WUT social network for different classes extracted based on the relationship strength

| Motif ID | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Class 12345 | 363.9 | –60.8 | –134.3 | –158.6 | 98.8 | 230.3 | –184.5 | 161.8 | 35.7 | –219.6 | 144.2 | 279.8 | 167.4 |
| Class 2345 | 210.2 | 2.0 | –234.4 | –105.8 | 143.5 | 418.7 | –236.4 | 180.1 | 40.3 | –189.5 | 180.8 | 360.8 | 237.4 |
| Class 345 | –131.6 | –58.6 | –496.5 | 155.6 | 372.7 | 533.3 | –630.0 | 233.0 | 57.4 | –256.9 | 404.8 | 589.5 | 651.1 |
| Class 45 | –223.8 | –109.7 | –464.4 | 288.9 | 400.3 | 487.7 | –972.3 | 196.9 | 35.3 | –269.5 | 427.7 | 606.4 | 1279.0 |
| Class 5 | –305.3 | –160.4 | –307.5 | 500.6 | 318.1 | 312.1 | –589.5 | 94.3 | 13.4 | –293.5 | 382.6 | 335.1 | 1754.3 |
| Class 4 | –150.6 | –207.3 | –220.9 | –61.9 | 358.2 | 247.5 | –228.9 | 257.1 | 154.6 | –219.1 | 229.5 | 235.1 | 541.1 |
| Class 3 | 6.3 | –86.0 | –36.0 | –96.4 | 122.2 | 54.2 | –29.5 | 55.5 | 65.0 | –47.6 | 45.5 | 36.0 | 90.4 |
| Class 2 | 87.4 | –77.4 | –29.5 | –76.7 | 68.4 | 46.8 | –44.1 | 52.6 | 44.6 | –46.9 | 32.0 | 35.4 | 56.6 |
| Class 1 | 50.7 | –32.2 | –35.6 | –42.0 | 32.2 | 37.3 | –26.4 | 13.2 | 10.6 | –14.5 | 9.4 | 14.1 | 90.8 |

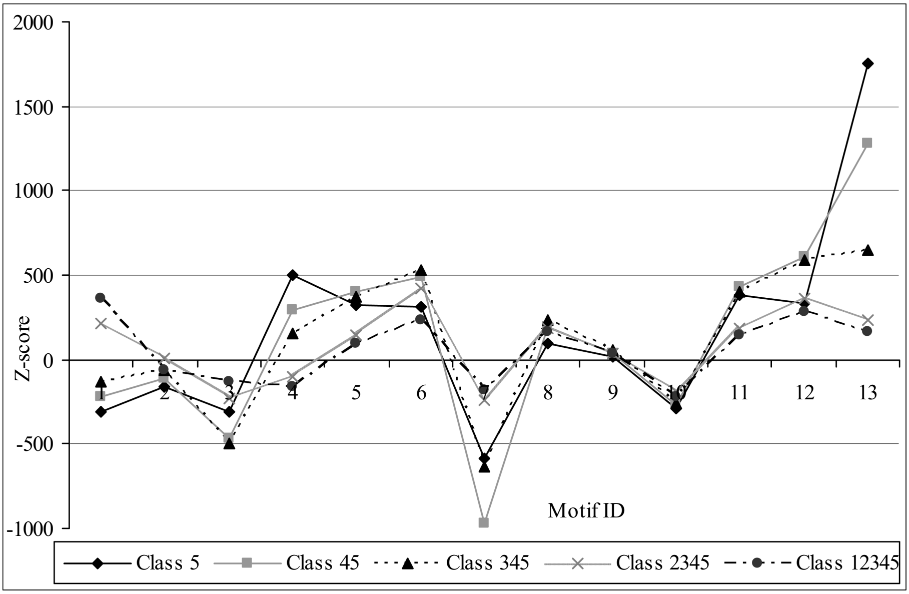


**Fig. 4.** Triad Significance Profile of the WUT email-based social network for different classes of relationship strength

characterized by intense communication (Table 2). We clearly see that the high positive Z-score of M13, big negative Z-score of M7 and decreasing with the growing strength of the ties Z-score of M1 may be called markers of social small-world networks and suggest increasing clustering level (Fig. 4). Also, the change for M1's Z-score from positive to negative values met when passing to subnetworks composed of stronger ties shows the fading of broadcast-type communication in strongly connected subnetworks. We also see that the subgraphs composed of the stronger ties are "more connected" and clustered then their lightly connected counterparts.

The above conclusions are additionally confirmed by TSPs for separate classes of the ties (Fig. 5). They reveal the growing importance of M13 and M7 for subnetworks of strong ties (Class 4 and 5) even in more convincing way. There is also one more observation concerning M4, which is dual-edge triad with two unidirectional edges pointing to the same node (Fig.2). Only for Class 5, we observed positive Z-score of M4, which is unusual for the investigated social network as a whole – compare with TSP of other classes in Fig. 4 and 5 as well as Table 2. This can be interpreted as above-statistical frequent occurrence of *receiver nodes* – we may treat them as of *executives*. They intensively receive reports while simultaneously being involved in dense clusters and frequent communication activities inside small-world network structures. Note that this effect is characteristic only for intensively communicating subnetwork and it has not been detected in the full communication graph, i.e. based on all classes of ties. It is also absent in TSPs of small social networks and WWW network studied in [9].

Overall, social networks created upon different classes and their various combinations belong to the same *superfamily* of networks [9] – their TSPs have the same shape regardless the range of relationship strengths *RS* (Fig. 3, 4, and 5). However, this general profile is most noticeable for stronger human relationships, especially within Class 5. In other words, Class 5 containing only 11% of all edges (Class 12345) but as much as over 20% of total mutual ties (Table 1) appears to be is the

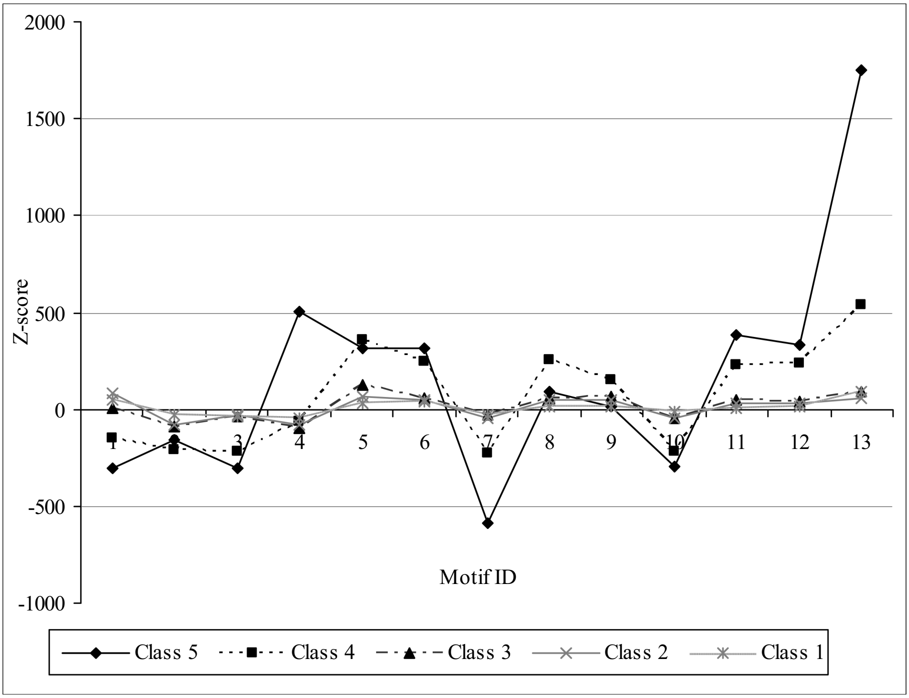


**Fig. 5.** Triad Significance Profile of the WUT social network for different classes of the relationship strength

most distinctive representative of the entire social network. Hence, Triad Significance Profile for strong human relationships can be treated as the small, condensed pattern that reflect the character of the entire social network.

## 4 Conclusions

Network motif analysis is the fast method to discover general profile of the entire network in the compact form since small motifs reflect patterns of the common local topology. In the email-based social networks, motifs preserved their distribution for all analyzed networks. Stronger relationships in the email-based social network are more mutual. Moreover, the motif-based fingerprint (TSP) is more distinctive for the network created from the stronger relationships. Besides, intensive communication results in greater frequency of more complex motifs and greater clustering level.

Further research will focus on multirelational social networks [10] as well as on dynamical properties of motif analysis in large social networks. New fast algorithms for motif detection may be applied to periodically gathered communication logs in order to discover time-related changes in large, evolving social structures.

**Acknowledgments.** The work was supported by The Polish Ministry of Science and Higher Education, grant no. N516 037 31/3708.

## References

1. Barabasi, A.-L., Albert, R.: Emergence of scaling in random networks. Science 286, 509–512 (1999)
2. Chung-Yuan, H., Chuen-Tsai, S., Chia-Ying, C., Ji-Lung, H.: Bridge and brick motifs in complex networks. Physica A 377, 340–350 (2007)
3. Itzkovitz, S., Milo, R., Kashtan, N., Ziv, G., Alon, U.: Subgraphs in random networks. Physical Review E 68, 026127 (2003)
4. Kashtan, N., Itzkovitz, S., Milo, R., Alon, U.: Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. Bioinformatics 20(11), 1746–1758 (2004)
5. Kazienko, P., Musiał, K., Zgrzywa, A.: Evaluation of Node Position Based on Email Communication. Control and Cybernetics (to appear, 2008)
6. Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., Alon, U.: Network motifs: simple building blocks of complex networks. Science 298, 824–827 (2002)
7. Mangan, S., Alon, U.: Structure and function of the feedforward loop network motif. Proc. of the National Academy of Science, USA 100(21), 11980–11985 (2003)
8. Mangan, S., Zaslaver, A., Alon, U.: The coherent feedforward loop serves as a sign sensitive delay element in transcription networks. J. Molecular Biology 334, 197–204 (2003)
9. Milo, R., Itzkovitz, S., Kashtan, N., Levitt, R., Shen-Orr, S., Ayzenshtat, I., Sheffer, M., Alon, U.: Superfamilies of evolved and designed networks. Science 303(5663), 1538–1542 (2004)
10. Musiał, K., Kazienko, P., Kajdanowicz, T.: Multirelational Social Networks in Multimedia Sharing Systems. In: Nguyen, N.T., Kołaczek, G., Gabryś, B. (eds.) Knowledge Processing and Reasoning for Information Society, ch. 18, pp. 275–292. EXIT, Warsaw (2008)

11. Shen-Orr, S., Milo, R., Mangan, S., Alon, U.: Network motifs in the transciptional regualtion network of Escherichia coli. Nat. Genet. 31, 64–68 (2002)
12. Valverde, S., Theraulaz, G., Gautrais, J., Fourcassie, V., Sole, R.V.: Self-organization patterns in wasp and open source communities. IEEE Intelligent Systems 21(2), 36–40 (2006)
13. Wernicke, S.: Efficient detection of network motifs. IEEE/ACM Transactions on Computational Biology and Bioinformatics 3(4), 347–359 (2006)
14. Wernicke, S., Rasche, F.: FANMOD: a tool for fast network motif detection. Bioinformatics 22(9), 1152–1153 (2006)

# A Semantic Language for Querying Anonymous Web Sources

François Pinet[1] and Michel Schneider[1,2]

[1] Cemagref, 24 Avenue des Landais, 63172 Aubière Cedex, France
[2] LIMOS, Complexe des Cézeaux, 63173 Aubière Cedex, France
{francois.pinet, michel.schneider}@cemagref.fr

**Abstract.** A great deal of work has been carried out in recent years to facilitate access to data and information available on the Web. Proposals converge in two additional areas which consist in providing the sources with semantic annotations and in designing languages and tools that are capable of using these annotations. However, a large number of sources have not yet been annotated suitably. Besides, languages and existing tools do not allow the user to formulate "blind" queries without knowing the sources. To overcome these two limitations, in this paper we propose a flexible query language which allows a user to query sources in an anonymous way without knowing their existence and their structure. Queries can be solved by a system which in advance discovers potential sources and memorizes their schemas. We clarify how such a system can function.

## 1 Introduction

A large number of works in recent years have shown interest in the problem of retrieving data, documents, and knowledge available on the Web. The need to annotate available resources in order to facilitate their search is now well recognized. Complementary proposals have therefore been formulated, on one hand to annotate resources in computable semantic forms, and on the other hand to implement languages and systems capable of exploiting these annotations. Particular efforts have been devoted to the retrieval of knowledge and meta-data by benefiting from W3C standards. A synthesis of this trend can be found in [1]. It can also be observed that a certain number of works use these same standards for the retrieval of texts in documents [17]. Data retrieval from structured or semi-structured sources has also been subject to numerous investigations notably with a view to integration and mediation [3, 6, 15, 23]. The advantage of this type of approach is that it offers the user an integrated view of the data through a global schema. Its main drawback is the difficulty in maintaining the global schema when a source evolves or when a new source is integrated. Blind interrogation (without a view of a schema) of separate sources is an approach which has been investigated only to a lesser extent. Nevertheless such needs do exist. A simple example is the search for bibliographical information about works and authors among the different bibliographical sources available on the Web (e.g. what is the price of the book "Introduction to relational Databases", one author of

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 106–116, 2008.

which is "Date"). It would be interesting to have a simple and structured language for querying these sources without knowing what their schema is (or even their existence), only from the concept names of the domain and from their relationships.

The works concerned by data searches in XML sources from keywords are quite close to our objective [2, 10, 14]. In these works it is a question of offering the user simple access without having to use Xpath or Xquery. Therefore an interface based on keywords is recommended. It is indeed a very simple device for the user but it introduces ambiguities in the specification of relationships (between concepts) and results. This problem of ambiguities is the source of interesting challenges for which satisfactory solutions are not yet completely available. To a certain extent these preoccupations are similar to those aiming to improve the efficiency of Web search engines [5, 7, 12, 13, 21, 22, 24].

We are convinced that it is possible to propose a simple and flexible language for the end user to achieve this kind of objective. The language SemanticSQL, which we present in this paper, interprets the intention of the user from the names of concepts and their relationships. It does not require the user to know the schemas of the sources or their existence. This language is qualified as flexible because it can adapt to badly annotated sources as well as richly annotated sources. We further show that a query in this language can be fairly matched with the schemas of relational sources or XML sources. We then sketch the functioning of a system capable of interpreting and resolving a user query. Such a system must discover the potential sources available on the Web in advance and memorize theirs schemas.

The paper is organized as follows. In section 2 we present the syntax of our language. In section 3 we give some indications on how the user and the system can co-operate. Section 4 defines the notion of valid query. In section 5 we explain the general matching process between a query and a potential source. Section 6 sets out details on how links matchings can be established. Section 7 provides comments on the results of certain experiments. Section 8 concludes and suggests some perspectives.

## 2 Syntax of the SemanticSQL Query Language

The syntax of our query language is organised in the same fashion as SQL. The query is divided up into three clauses with the reserved words "select", "from" and "where". As in SQL the "select" clause specifies the result of the query and the "where" clause specifies the condition which must be satisfied. The "from" clause has a different meaning. Since in our approach the user does not know the schemas of the sources, it cannot designate the structures from which the result must be extracted. So in the "from" clause the user indicates the names of types which he thinks might exist in the sources. An optional "on" clause makes it possible to restrict the query to certain sources.

The general form of a query is as follows:

**Select** list_of_variables **from** list_of_type_variable_pairs
[**where** list_of_conditions] [**on** list_of_sources]

list_of_variables: This is a list of variables introduced by the user. Names of variables are freely chosen by the user and are separated by commas. Names of variables are not related to the semantics of the query type_variable_pair: This is a pair separated by a space composed of a type name and a variable name. It is used to allocate a variable to a type, so a variable marks an instance of a given type. Names of types in this clause are of great importance. They are directly related to the semantics of the query. Such pairs allow the user to specify the meaning of the objects which he is looking for.

list_of_type_variable_pairs: This is a list of type_variable_pairs separated by commas.

list_of_conditions: This is a list of conditions separated by commas where the comma has a particular significance: it marks the "and" operator. So the global condition of the where "clause" is true if each separate condition in this list is true.

condition := valuation_condition | link_condition

valuation_condition := elementary_condition |
(valuation_condition_1 or valuation_condition_2) |
(valuation_condition_1 and valuation_condition_2) |

elementary_condition := variable_name op constant | variable_name_1 op variable_name_2

op := > | < | >= | <= | <> | =

link_condition: := variable_name_1()variable_name_2 |
variable_name_1(role_name)variable_name_2 |
variable_name_1@variable_name_2

The first kind of conditions is used to constrain the values of a variable.

The second kind of conditions allows the user to specify the existence of links between query variables. This specification is made on a pair basis with the double symbol (). For example a()u means that the two variables a and u must be connected through some link. We can indicate the meaning of the link between the parentheses. For example a(write)b means that a must be connected to b through a link which means that "a writes b". The connection between a type and one of its attributes or data type properties is specified by the notation a@n.

The notation a()b specifies a non oriented link. The other two notations correspond to oriented links (from variable1 to variable2).

list_of_sources: This is a list of source names separated by commas.

The "where" and the "on" clauses are optional.

*Example*: Here is an example of a concrete query

Q1: Select n, t from book b, title t, author a, name n, university u where a(write)b, a@n, b@t, a()u, u='Stanford'

Query Q1 can be paraphrased as follows: search for the tuples [n, t] where n is an attribute which represents the name of an author a, t is an attribute which represents the title of a book b, a wrote b, a is linked to a university, the value of which is 'Stanford' (the semantics of the link does not matter).

*Interpretation of a query*: The interpretation of a query in our language is as follows: give all possible values for the list_of_variables which satisfy the condition of the where clause, each variable taking its values from among those of the type it represents.

## 3 Respective Responsibilities of the User and the System

In our approach we suppose that the user possesses some knowledge about the domain and its potential information (names of types, structures of these types). But it is not required that he knows the schema of the sources. In this section we give some indications on how the user and the system can cooperate in order to pose and to answer a query.

*Source names*
The user does not need to know the source names or the existence of certain sources. In such a case, the system should choose those that are the most suited to the user query from among the sources of its directory. Further along the user can indicate to the system sources concerned by its query.

*Type names*
The user chooses the most suitable names for each of the types concerned by his query. The choice of these type names is particularly important because it induces the interpretation of the query by the system. The user may be able to use an ontology of the domain known from the system. The system can widen its interpretation to synonyms and even to hyponyms of each of the names.

*Link conditions*
Even then, the user will fill in the links that he considers to be most plausible between the types which he has specified in the "from" clause. A named link should be preferred in order to facilitate the resolution of the query. However names of links (which correspond to names of properties or associations in the schemas of certain sources) can show considerable variety. This variety can make it difficult for the system to identify the sources which correspond best to the user's wish. Moreover links are not always named (as in the case of relational and XML sources). In certain cases it will be better to choose an anonymous link and to let the system identify the possibilities offered with sources. If the number of possibilities is too great, a dialogue with the user can make it possible to limit the search space. Usage of links such a@b can be restrictive. Certain sources (notably XML sources) can treat attributes as standard elements. So it will be necessary for the system to look for all the possible links which can exist between a and b even though the user indicates an attribute.

*Valuation conditions*
By specifying a valuation condition of the form "variable_name_1 op constant" the user supposes that type T associated to variable_1 is atomic and is compatible with the type of the constant. If for a given source, this type T is not atomic, a correct comparison cannot occur. So the system may search, from among the attributes or the descendants of this type, for some atomic type which can represent type T and whose value can be compared with the constant. This atomic type must be representative of type T. We can use the ontology to help in its identification. For example a non atomic type can be represented by an atomic type whose name is: "name", "label", …. The same problem arises for a valuation condition of the form "variable_name_1 op variable_name_2". For each type associated to the two variables, atomic descendants will need to be looked for whose values can be compared.

*Form of the result*

Depending on the sources, a type in the result can be atomic or structured. The display of the value of a structured type can create problems. It can be agreed that the value is the concatenation of the values of its atomic types. Alternatively, the system can display only the value of an element which identifies the result.

## 4 Valid Query

In order to be valid, a query has to respect certain constraints. We shall clarify these constraints by representing the query by a graph.

A query graph is constructed as follows. Each type is represented with a node which carries the name of the type. Each link between two variables is represented with an edge between the two associated types. A link a()b is represented with a non directed edge. A link a(role) b is represented with a directed edge from a towards b; this edge is labelled by role. A link a@b is represented with a directed edge in a dotted line from a towards b.

*Definition (valid query)*: A query is valid if its graph is connected and with at most a single edge between each pair of nodes.

These constraints are justified as follows. First, if the graph is not connected, there are at least two separated components which each correspond to independent sub-queries. If there are two edges between the same pair of nodes, then there is a redundancy or incoherence in the meaning of the link which the user wishes to specify between the corresponding types.

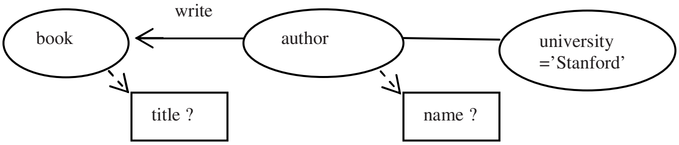


**Fig. 1.** Query graph of Q1

Query graph of Q1 is represented in figure 1 (the types are represented by ovals and the attributes by rectangles). This graph respects the two validity constraints and query Q1 is therefore valid.

## 5 General Principles for Matching a Query with Potential Sources

Answering the query means finding a correspondence between the query and each of the schemas of the potential sources.

Since a query can be represented by a graph, a solution for this problem consists in using a matching technique to establish this correspondence. Many matching algorithms have been suggested [4, 8, 16, 19, 20]. However these algorithms are difficult to adjust. Besides, a query in our language is a very simple object compared to a schema and it would be interesting to study specific approaches for establishing a matching.

In this section we also provide a number of general indications on how the system can answer a query expressed with our language.

The matchinghas to consider two kinds of element: the names of types and the links between types. To obtain suitable answers for a query we suggest an approach consisting in first establishing the matching between the type names. It is only when a matching can be found for each type name that a source will be selected for the matching of links.

The matching of names can be very difficult if names are constructed freely by the users and the designers. We will suppose, in order to facilitate this stage, that the names of types in queries respect a domain ontology. So the ontology can provide a list of synonyms and a list of hyponyms for each type name of the query. The matching of a name with a source is then tested first with the name itself, then with its synonyms, then with its hyponyms. We do not require the names in the sources to respect the ontology. So names in sources are previously transformed in order to facilitate matching. Different kinds of string transformations have been proposed [4] and some of them are very efficient. For a given pair (a query, a schema), several possibilities of matching can exist for each query name. It is necessary to consider all the possibilities for the following stage of matching the links.

## 6 Some Guidelines for Matching the Links with XML Sources

In this section we provide a certain number of guidelines for matching links between a query and the schema of a XML source. Similar guidelines can be drawn with relational sources.

To illustrate these guidelines we consider the XML source of figure 2 which is a potential candidate for answering Q1.

The matching of type names gives the following correspondences:

title → title (of book)
name → name (of author), name (of grading_university, ofworking_university)
book → book
author → author
university → grading_university, working_university

So each type name for the query has at most one correspondence in the source. Subsequently we can try to match the links.

For the links book()title and author()name we can detect the correspondents easily (link between an element and one of its son for the first, link between an element and one of its attribute for the second). For the link author(write)book, it does not appear in the source a direct link between "author" and "book". But we observe that the word "writing" matches with "write" (the correspondence is established after lemmatization). So, we can compose the author→writing link (child-parent link) with the writing→book link (parent-child link) to establish a correspondence. Another problem is that there are two links between "author" and "university" with two different meanings. In this case the system must propose the two possibilities and the user can then choose one of them. The last problem is that of the valuation condition university = "Stanford”. If we consider the first correspondence (university→ grading_university), it is not possible to make the comparison directly. The system must detect that there is an atomic type which can represent the grading_university type and which permits the comparison. The attribute "name" can play this role. The automatic detection of such a possibility in all the possible situations is a difficult problem. The ontology of the domain can provide a certain number of indications. Otherwise, the system can display a partial schema of the source in order to ask for the user's opinion. For this valuation condition, another alternative is possible by considering the other correspondence university→ working_university. In this other case the comparison can be made using also the attribute "name".

A matching can thus be established for each link. The solution is not unique because there are two variants for the valuation condition.

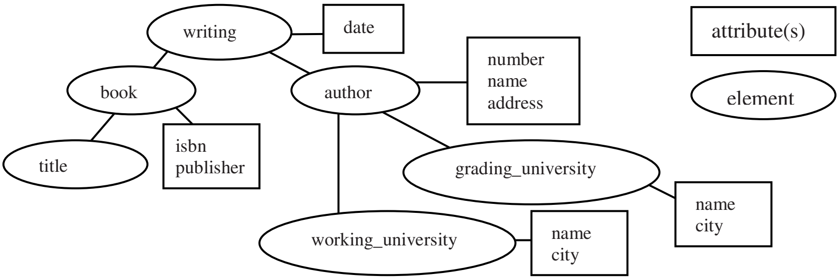


**Fig. 2.** XML tree for a source which matches with query Q1

In figure 3, we show the XML tree for another XML source with the same elements but organized differently. Although the link between book and author is not semantically characterized, the system must select this source for a matching with query Q1. The user will then confirm or not.

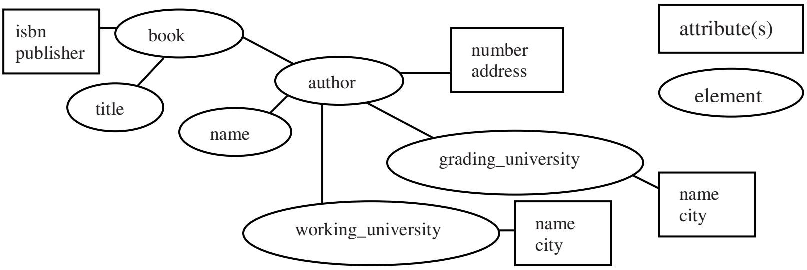


**Fig. 3.** XML tree for a second source which matches with query Q1

The XML source for which the element book would be a child of the element author (and not the parent as in figure 4) must also be selected by the system for the same query.

We formalize the previous analysis by the following rules.

R1: The link a()b of the query matches with a source S if A is a correspondent of a in the source, B is a correspondent of b and one of the following conditions holds:

i) it exists a direct link A,B or B,A in the source or B is an attribute of A or A is an attribute of B
ii) there exist a direct link A,C or C,A in the source and a path C,…,B where C and each intermediate node are in a relation of synonymy or hyperonymy or hyponymy with B
iii) same as ii) by exchanging the roles of A and B.

R2: The link a(c)b of the query matches with a source S if A, B, C are respectively correspondents of a, b, c in the source and one of the following conditions holds:

i) there exist a direct link A,C or C,A and a direct link C,B in the source or (symmetric case) there exist a direct link B,C or C,B and a direct link C,A
ii) there exist a path A,…,C or C,…,A in the source and a path C,…, B where each intermediate node is in a relation of synonymy or hyperonymy or hyponymy with A (resp. B)
iii) same as ii) by exchanging the roles of A and B.

R3: The link a@b of the query matches with a sources S if A is a correspondent of a in the source, B is a correspondent of b and one of the following conditions holds:

i) B is an attribute of A or a simple child of A
ii) it exists a path A,…,B in the source where each intermediate node is in a relation of synonymy or hyperonymy or hyponymy with A and B is a simple element
iii) it exists a path A,…,E in the source where each intermediate node and E are in a relation of synonymy or hyperonymy or hyponymy with A and B is an attribute of E.

R4: The condition a=v of the query matches with a source S if A is a correspondent of a in the source and one of the following conditions holds:

i) A is a simple element or an attribute (the condition is tested with A)
ii) it exists a path A,…,D in the source where each node is in a relation of synonymy or hyperonymy or hyponymy with A and D is a simple element (the condition is tested with D)
iii) it exists a path A,…,E where each intermediate node and E are in a relation of synonymy or hyperonymy or hyponymy with A and it exists an attribute of E having a name like ("name", "label", "description") (the condition is tested with this attribute.

Similar rules can be specified for relational sources.

## 7 Prototype and Experiments

To verify the ability of the language and the feasibility of our approach we have built a small prototype and conducted a number of experiments with XML sources. We used WORDNET [18] to help with the matching of names. Access to WORDNET was made through the JAVA API Java WordNet Library [11]. The body of the matcher was written in JAVA. For every link of the query, the various cases of matchings identified in section 5 are tested successively by considering all the possibilities offered with synonyms and hyponyms (at most level 3) for names. However matching of names is restricted to the domain (by using the ontology) to avoid meaning misunderstanding. We have only tested queries with links of the type a()b and conditions.

In the first stage, our experiments were conducted on six different XML sources containing data on sales of products. Sources were built manually. They contained from 8 to 14 elements. Every element had on average two attributes. We submitted ten different queries to the system. The matchings performed quite well. 90% of the total returned matchings were correct and 85% of the total correct matchings were retrieved. The incorrect matchings resulted essentially from a bad detection of the type replacement in the valuation conditions.

In the second stage we implemented a module in our prototype to discover potential sources in the same domain (sales of product) on the Web and to memorize their DTD. After validation by the administrator, 10 different sources were then incorporated into the system. We submitted the same 10 queries and we observed that the matchings performed badly. It appears in fact that several element names in the DTD were abbreviations which cannot be handled correctly by our string transformations. So we decided to create a specific dictionary for the management of these abbreviations and we significantly increased the efficiency of the matches. 80% of the returned matchings were correct. Bad type replacements partially explained incorrect matchings. Another cause of error was observed: a nonsense in the matching of names. These sources were rather complex and it was difficult to list manually all the correct matchings for each query. We estimated that the prototype had discovered about 75 % of them.

From these experiments it appears that the approach is realistic. The main points which condition its efficiency are the type replacement in valuation conditions and the detection of the meaning of names.

## 8 Conclusion and Perspectives

In this paper we proposed a query language for a final user which allows blind accesses to Web data sources. The user formulates his query by specifying the names of types and relationships between these types. It is not necessary for the user to know the existence of sources or their schemas.

This language is flexible. It can adapt to sources whether well annotated or not.

We discussed how a system can analyze a query and elaborate the results. Such a system must discover the potential sources in advance (for the considered domain) and memorize their schemas. To answer a query it first has to look for the matchings of names and then for the matchings of links. It proposes all the solutions and the user validates those with which he is interested. The system can then rewrite the query for the corresponding sources.

We conducted experiments with XML sources which establish the efficiency of the language and the feasibility of the approach. Matchings can be improved on two points: type replacement in the valuation conditions and detection of the meaning of names. Concerning this last point one can reduce the ambiguities by working in a precise domain. It is possible also to take into account the profile of the user which can be acquired through a dialogue or by observing the queries. One can also observe how the user validates the solutions proposed by the system. Another way is to group results according to their different meanings [9].

Others improvements can be envisaged. One could exploit the mappings which can exist between sources to permit joins between these sources. The graph of a query is very simple and expressive and it would be interesting to study how it can support a graphical query interface. To improve the matching of links one also can ask the user to indicate the cardinalities.

Such a system can be very useful for different applications. Incorporated into an intranet system, it would allow a user to reach the data sources without knowing their schemas, by being based only on the domain ontology. In a P2P system, it could be installed on some peers or on super-peers to facilitate access to data by their semantics.

## References

1. Bailey, J., Bry, F., Furche, T., Schaffert, S.: Web and Semantic Web Query Languages: A Survey. Reasoning Web, 35–133 (2005)
2. Cohen, S., Mamou, J., Kanza, Y., Sagiv, Y.: XSEarch: A Semantic Search Engine for XML. In: VLDB 2003, pp. 45–56 (2003)
3. Cui, Z., Jones, D., O'Brien, P.: Issues in Ontology-based Information Integration. In: IJCAI, Seattle (2001)
4. Do, H.H., Rahm, E.: COMA - A System for Flexible Combination of Schema Matching Approaches. In: VLDB 2002, pp. 610–621 (2002)
5. Duke, A., Glover, T., Davies, J.: Squirrel: An Advanced Semantic Search and Browse Facility. In: Franconi, E., Kifer, M., May, W. (eds.) ESWC 2007. LNCS, vol. 4519, pp. 341–355. Springer, Heidelberg (2007)
6. Garcia-Molina, H., Papakonstantinou, Y., Quass, D., Rajaraman, A., Sagiv, Y., Ullman, J., Vassalos, V., Widom, J.: The Tsimmis approach to mediation: Data models and languages. Journal of Intelligent Information Systems 8(2), 117–132 (1997)
7. Goldschmidt, D.E., Krishnamoorthy, M.S.: Comparing keyword search to semantic search: a case study in solving crossword puzzles using the Google$^{TM}$ API. Softw. Pract. Exper. (SPE) 38(4), 417–445 (2008)
8. Hai Do, H., Melnik, S., Rahm, E.: Comparison of Schema Matching Evaluations. Web, Web Services, and Database Systems, pp. 221–237 (2002)
9. Hemayati, R., Meng, W., Yu, C.T.: Semantic-Based Grouping of Search Engine Results Using WordNet. In: Dong, G., Lin, X., Wang, W., Yang, Y., Yu, J.X. (eds.) APWeb/WAIM 2007. LNCS, vol. 4505, pp. 678–686. Springer, Heidelberg (2007)
10. Hristidis, V., Koudas, N., Papakonstantinou, Y., Srivastava, D.: Keyword Proximity Search in XML Trees. IEEE Trans. Knowl. Data Eng (TKDE) 18(4), 525–539 (2006)

11. JWNL. Java WordNet Library, `http://sourceforge.net/projects/jwordnet`
12. Kandogan, E., Krishnamurthy, R., Raghavan, S., Vaithyanathan, S., Zhu, H.: Avatar semantic search: a database approach to information retrieval. In: SIGMOD 2006, pp. 790–792 (2006)
13. Li, Y., Wang, Y., Huang, X.: A Relation-Based Search Engine in Semantic Web. IEEE Trans. Knowl. Data Eng (TKDE) 19(2), 273–282 (2007)
14. Liu, Z., Walker, J., Chen, Y.: XSeek: A Semantic XML Search Engine Using Keywords. In: VLDB 2007, pp. 1330–1333 (2007)
15. Lenzerini, M.: Logical Foundations for Data Integration. In: Vojtáš, P., Bieliková, M., Charron-Bost, B., Sýkora, O. (eds.) SOFSEM 2005. LNCS, vol. 3381, pp. 38–40. Springer, Heidelberg (2005)
16. Madhavan, J., Bernstein, P.A., Rahm, R.: Generic Schema Matching with Cupid. In: VLDB 2001, pp. 49–58 (2001)
17. Mangold, C.: A survey and classification of semantic search approaches. IJMSO 2(1), 23–34 (2007)
18. Miller, G.: Wordnet: A Lexical Database for English. Communications of the ACM 38, 39–41 (1995)
19. Mohsenzadeh, M., Shams, F.: Teshnehlab M.: Comparison of Schema Matching Systems. WEC (2), 141–147 (2005)
20. Rahm, E., Bernstein, P.A.: A survey of approaches to automatic schema matching. VLDB Journal 10(4), 334–350 (2001)
21. Royo, J.A., Mena, E., Bernad, J., Illarramendi, A.: Searching the Web: From Keywords to Semantic Queries. In: ICITA 2005, pp. 244–249 (2005)
22. Tran, T., Cimiano, P., Rudolph, S., Studer, R.: Ontology-Based Interpretation of Keywords for Semantic Search. In: Aberer, K., Choi, K.-S., Noy, N., Allemang, D., Lee, K.-I., Nixon, L., Golbeck, J., Mika, P., Maynard, D., Mizoguchi, R., Schreiber, G., Cudré-Mauroux, P. (eds.) ASWC 2007 and ISWC 2007. LNCS, vol. 4825, pp. 523–536. Springer, Heidelberg (2007)
23. Wiederhold, G.: Mediators in the architecture of future information systems. IEEE Computer 25(3), 38–49 (1992)
24. Zhou, Q., Wang, C., Xiong, M., Wang, H., Yu, Y.: SPARK: Adapting Keyword Query to Semantic Search. In: Aberer, K., Choi, K.-S., Noy, N., Allemang, D., Lee, K.-I., Nixon, L., Golbeck, J., Mika, P., Maynard, D., Mizoguchi, R., Schreiber, G., Cudré-Mauroux, P. (eds.) ASWC 2007 and ISWC 2007. LNCS, vol. 4825, pp. 694–707. Springer, Heidelberg (2007)

# CDNs with Global Adaptive Request Distribution*

Leszek Borzemski[1] and Krzysztof Zatwarnicki[2]

[1] Institute of Information Science and Engineering,
Wroclaw University of Technology, Wroclaw, Poland
Leszek.Borzemski@pwr.wroc.pl
[2] Faculty of Electrical Engineering, Automatic Control and Computer Science,
Opole University of Technology, Opole, Poland
KZatwarnicki@po.opole.pl

**Abstract.** This paper presents the application of fuzzy logic and neural networks to HTTP request dispatching performed within Content Delivery Network. We propose a global request distribution algorithm called GARD to support request routing to the surrogate servers that deliver the requested content in an efficient manner. The algorithm uses the fuzzy-neural decision-making mechanism to assign each incoming request to the server with the least expected response time. The response time include the transmission time over the network, both for the request and for the response, as well as the time elapsed on the server responding to the request. We demonstrate through the simulations that our algorithm is more effective than popular dispatching policies as Round-Robin and Weighted Round-Robin. We also show that in case of non-evenly loaded environment the GARD algorithm outperforms RTT algorithm which is often used in CDNs.

**Keywords:** CDN, Dispatching algorithms, Fuzzy-neural HTTP request distribution, Web service quality, Distributed systems.

## 1 Introduction

Content Delivery Networks (CDNs) are modern Internet services replicating contents over several Web servers called surrogate servers. Such servers are placed at various geographical locations close or at the network edges in order to improve performance and scalability in access to Web resources. CDNs improve network performance by maximizing bandwidth and improving accessibility. Typically, they are used by Internet service providers as well as global companies and enterprises.

Generally, in CDN-like environment, a Web user is served with the content from the nearby replicated Web server and retrieves resources from that server instead the original server specified its URL in the Web browser. Using the proprietary algorithm, the CDN network reroutes the requests to embedded objects to replica servers which are ‘closest’ to the client, in order to serve them as quick as possible.

* This work was supported by the Polish Ministry of Science and Higher Education under Grant No. N516 032 31/3359 (2006-2009).

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 117–124, 2008.

There are various existing commercial CDNs (e.g., Akamai, Mirror Image, Netli, Value CDN) as well as academic CDNs (e.g., Codeen, COMODIN, Coral, Globule). Current research and exploitation experience show the need to improve functionality and performance of a request-routing mechanism which is a basis of CDN functioning [10]. This mechanism called also as the *distribution* algorithm is responsible for routing client requests to an appropriate surrogate server for the delivery of content. It directs client requests to the replica server 'closest' to the client. However, the closest server may not be the best surrogate server for responding to the client request. Therefore, a request-routing mechanism should use more information and in more sophisticated way to select the ultimate server. It may use a set of metrics such as network proximity, client perceived latency, distance, and replica server load in an attempt to redirect user requests.

Neural and fuzzy decision making scheme has many significant applications in the deployment of intelligent computer systems [8, 9]. For example, fuzzy logic controllers can efficiently deal with the uncertainties and with intrinsic nonlinearities of controlled systems [8]. By combining neural networks and fuzzy systems we can get learning and adaptation characteristics of the controllers. We have used this approach to develop FNRD algorithm for fuzzy-neural adaptive HTTP request distribution in a local Web cluster system [3]. FNRD policy has been useful in the design of the content-aware Web switch for local Web sites [3, 4]. It outperforms other competitive dispatching policies including the state-of-the-art content-aware "reference" algorithms like CAP [7], LARD [12], and RR [6]. We have also developed the broker-based request global routing algorithm GARDiB for distributed Web systems to service HTTP requests in a time-effective manner [2].

The main contribution of this paper is to develop a new distribution algorithm called Global Adaptive Request Distribution (GARD) which uses a fuzzy-neural based decision model to optimize request response time in CDN systems.

The reminder of the paper is organized as follows. Section 2 presents the architecture of GARD based CDN system. Section 3 discusses the simulation model and the results of simulation experiments. Section 4 concludes our study.

## 2 Architecture of GARD Based CDN System

The GARD system is designed for the supporting of the service for the distribution of the content in a wide-area Internet network. GARD distributes the content using the HTTP network protocol and own global adaptive request distribution algorithm. Fig. 1 shows the general scheme of the GARD based content delivery network.

The system consists of three main components: clients, local services and content distribution servers. An *individual client* is the source of HTTP requests being sent for Web resources to explore *the content* of Web. The clients are equipped with their computers and appropriate software for Web resource browsing such as Web browsers. The whole content of the Internet is build from local contents presented by local Web sites as *local services*. Local services (LSs) are geographically distributed in a wide-area Internet network. They can be developed as locally clustered Web systems with fully replicated local content or as individual Web servers. In this paper we assume that all local services can respond to all client requests. Content distribution

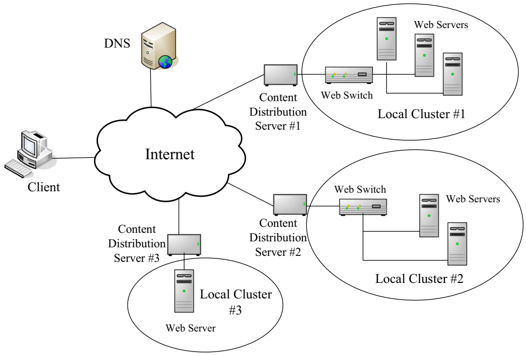


**Fig. 1.** Architecture of GARD based CDN system

servers (CDS) are front-end network appliances localized nearby of local services responsible for the content distribution in the replay of HTTP requests. Each local service is supported by its own CDS having an individual IP address.

Every client wanting to take a Web resource from a given Web site must to obtain the IP address of that Web site on the basis of Web site's mnemonic name (e.g. www.opole.pl). This is supported by the specific network service/server called DNS (Domain Name Service/Server). In our system the DNS transforms given mnemonic name into the IP address of the content distribution server being in the closest geographical distance from client's IP. Further client communication in Web resource gaining is made via that CDS server.

When the client sends its first HTTP request to the chosen CDS server, the connection time is measured by that server. Next, the CDS server redirects this request to the local service which is supported by that CDS server. Local service returns the specific Web resource. If the resource it returns is the HTML file, then the CDS server modifies this file by changing the addresses of embedded objects from local addresses (that is pointing at the content stored in that local service) to remote ones (that is pointing at the content stored at any other Web site) in such a way to include addresses from as many remaining local services as possible. As a consequence, a single object is sent from every local service. The rest of objects are gained from that local service which was accessed at the very beginning. The CDS server returns the HTTP response with the requested object and a demand to set a specific *cookie* value (LS identifier and the connection time) using a cookie mechanism. When client sends the requests to remaining local services to get objects embedded in that Web page, every individual CDS server writes its connection time in the cookie field value.

When the client accesses the CDS server next time to get the next HTML file, then CDS server modifies HTML file in such a way to replace all embedded objects links with links to those local services which are able to deliver requested objects in a timely effective manner, in the shortest time as perceived by the client. CDS server estimates the delivery times of requested HTTP objects to by clients on the basis of the knowledge about the model of local service, client-to-CDS server connection times as well as the knowledge about loading of remaining local services and the type of requested object. CDS servers exchange with information about local service loadings and about the knowledge needed to estimate times of request servicing.

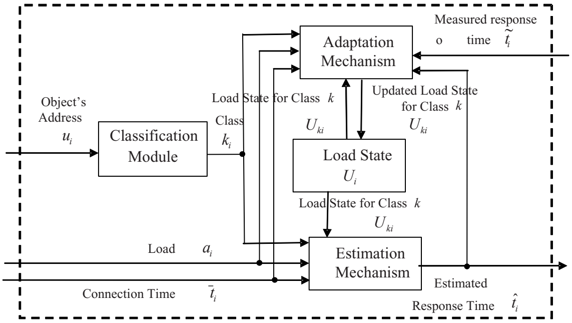


**Fig. 2.** Local service model

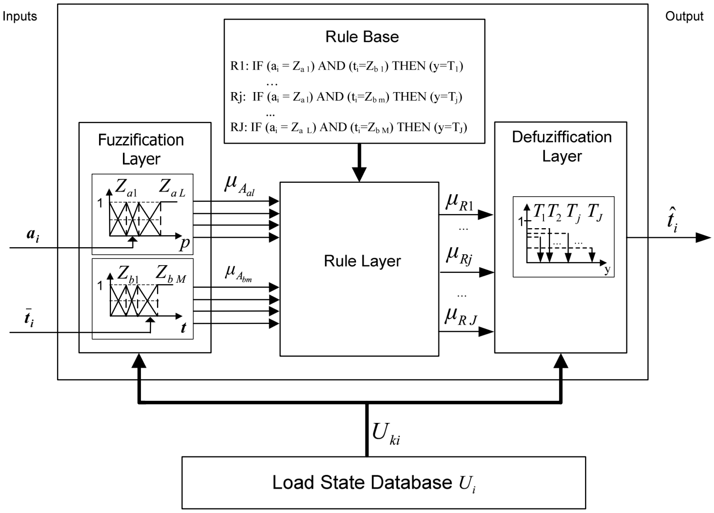


**Fig. 3.** LSM architecture based on Mamdami's model

Each request can be classified to one of classes; e.g. the $k$-th class, $k$=1,2,…, $K$. Local service model (LSM) takes into account the class of the request and load $a_i$ of local service for which the request response time is estimated ($a_i$ is the total number of request under service – this is the most current value we have got when request $i$ arrives), and the connection time $\bar{t}_i$ of the client to LS - this value is taken from cookie). The values of $\hat{t}_i^s$ for all sites are compared and the CDS chooses the LS with the minimum value of the estimated request response time. The address of chosen LS $u_i'$ replaces in HTML document the original address $u_i$ of particular object. Hence, the object is loaded from $u_i'$ address. LSM has four main components: *Classification Mechanism* (CM), *Estimation Mechanism* (EM), *Load State Database* (LSD), and

*Adaptation Mechanism* (AM). CM recognizes the class $k_i$ of object from $u_i$ address on the basis of object size and its category; $k \in \{1,...,K\}$ – we consider both static and dynamic objects. EM is built on the basis of the Mamdami's model [9] as shown in Fig. 3, AM is a neural component – both components combine into a neuro-fuzzy request distribution solution. The estimation system has three layers: *Fuzzification Layer*, *Rule Layer* and *Defuzzification Layer*. The EM determines the estimated request response time for object from $u_i$ address based on the $k_i$, $a_i$, $\bar{t}_i$ and $U_{ki}$ from $U_i = [U_{1i},...,U_{ki},...,U_{Ki}]$ LSD database which contains up-to-date information about the load state of the LS – this information is exchanged by local services. Based on the measured actual request response time $\tilde{t}_i$ the AM updates $U_i$ database in such a way that $U_{ki}$ for the $k_i$-th class is updated to a new value $U_{k(i+1)}$ - (this is for CDS that supports the LS which services $u_i'$).

## 3 Simulation Model and Experiment Results

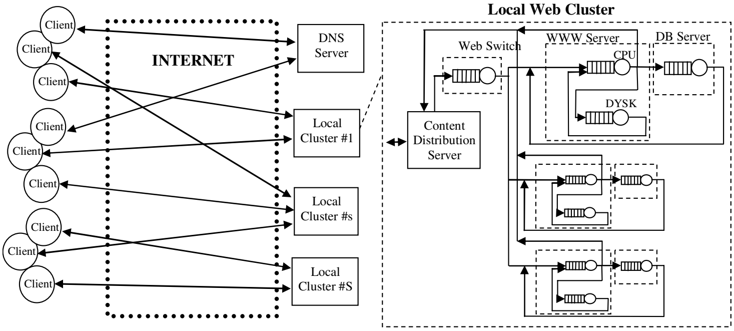
In the simulation experiments we develop a general model of global distributed Web system with local services made of clusters of Web servers (Fig. 4). We evaluated the performance using the workload model incorporating the most recent research on Web load. Local service was modeled as the Web cluster made of the Web switch employing FNRD local distribution algorithm [4], a number of WWW and database servers.

Internet module of the simulator was used to model the latency observed in Internet. The data transmission time for sending the request and receiving the response to the request is calculated as follows [10]:

$$data\_transmission\_time = RTT + \frac{object\_size + HTTP\_response\_header\_size}{throughput}$$

The Round-Trip Time (RTT) tells us how long a packet goes in the network from a source system to target system and back again. Object_size is the object (resource) length. The HTTP_response_header_size is usually 290 bytes. The throughput is determined as the effective number of bytes that can be transferred per second.

The distribution of RTT was modeled on the basis of the real-life end-to-end network measurements performed by the authors and targeting real-life Web sites in the Netherlands, Australia and [1]. The same Web document was downloaded to Opole, Poland network localization over a given time duration. We decided to choose the RFC1823 text file as it is often found at non-commercial sites that have non-overloaded Web servers. This file has a size of 47 KB which is not too small to measure how multiple IP packet file can be sent through the Internet and not too big to overload the target Web sites. Our Web active measurements were focused on the continuous observation of given Web sites performance characteristics through periodical measurement of request latency and transfer rate over a 48-hour period every 10 sec. From these experiments we obtained fifty various behaviors of Internet user a)



b) Static requests

| Category | Distribution | Parameters |
|---|---|---|
| Requests per session | Inverse Gaussian | μ=3.86, λ=9.46 |
| User think time | Pareto | α=1.4, k=1 |
| Objects per page | Pareto | α=1.33, k=2 |
| HTML object size | Lognormal<br>Pareto | μ=7.630, σ=1.001<br>α=1, k=10240 |
| Embedded object size | Lognormal | μ=8.215, σ=1.46 |

Dynamic requests (database)

| Type | Mean service time [msec] | Frequency |
|---|---|---|
| High intensive | 5 | 0.85 |
| Medium intensive | 10 | 0.14 |
| Low intensive | 20 | 0.01 |

**Fig. 4.** (a) A simulation model; (b) Workload model parameters

which were used in simulations. Each simulated client had a behavior randomly chosen from these fifty behaviors. The number of behaviors for each client was the same as the number of clusters in the system modeled.

Dynamic and static request have been considered. Static ones are served directly and only by the WWW servers, with the aid of the cache memory, whereas dynamic ones by database servers. We assumed a workload scenario consisting of 20% of dynamic requests for low, medium and high intensive workload sizes. The rest are static requests. We simulated three homogeneous LSs, each equipped with three Web and three database servers. The GARD algorithm proposed in this paper was compared with two popular request distribution algorithms: RR (Round-Robin) and WRR (Weighted Round-Robin) as well as with RTT algorithm deployed by Cisco. RR distributes the requests evenly among local services regardless of their loads. WRR takes into account that local services can serve requests with the different performance. In RTT algorithm when authoritative DNS server receives the request it sends the probes to DNS servers localized closely to local services. All probed servers respond at the same moment to the probe and that response that as the first one achieves the authoritative DNS server shows the IP of local service where the request is sent finally. The Content Distribution Server module shown in Fig. 4 simulates GARD activity as described in the section 2.

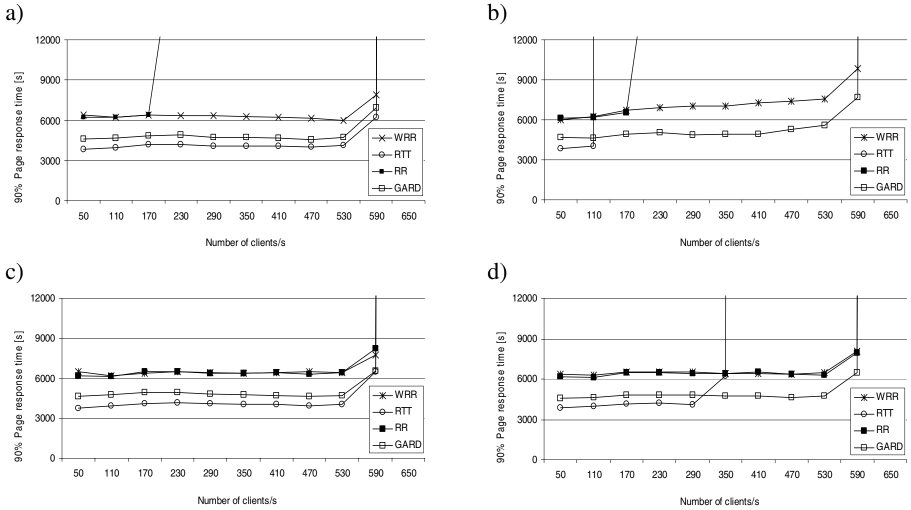


**Fig. 5.** 90-percentile of page response time vs. number of clients per second: a) 1/11%, 3/33%, 5/55% run, b) 1/55%, 3/33%, 5/11% run, c) 3/33%, 3/33%, 3/33% run, d) 3/11%, 3/33%, 3/55% run

In the first experiment we simulated the global system with three local services with one, three and five WWW servers. They were loaded proportionally to the number of WWW servers available in the service. Therefore we obtained the following shares of the total load: first service – 11% of total load, second one – 33%, and the last one – 55%. Such local service configuration and total load sharing is marked by the "1/11%, 3/33%, 5/55%" experiment sequence. The results of the experiment are shown in Fig. 5a. Fig. 5b shows the results for "1/55%, 3/33%, 5/11%" experiment sequence – the same clusters as in the previous experiment but an opposite load sharing. Fig. 5c presents the results for "3/33%, 3/33%, 3/33%" sequence – 3 clusters with 3 server and evenly loaded. Fig. 4d demonstrates the results of "3/11%, 3/33%, 3/55%" experiment where we have three the same clusters laded as in the first experiment. The main answer we wanted to obtain is how well RTT algorithm operates. It overcomes both RR and WRR, what we expected. However the results also show that RTT algorithm is a good one but only when the clients are balanced evenly and their number is directly proportional to the local service performance. Such situation is probably unreal in a real-life. The simulations also showed the proposed GARD algorithm overcomes all others for heavy loading in non-balanced environments (Fig. 5b and 5d).

## 4 Conclusion

We presented a novel fuzzy-neural HTTP global distribution policy called GARD which can be used in Content Delivery Networks for routing user requests to the server delivering the content. It achieved very good performance for the whole load range. Our policy outperformed very popular RR and WRR algorithms. The simulations showed that our algorithm shows better performance than well known in CDN systems RTT algorithm in case of non-evenly loaded environments.

## References

1. Borzemski, L., Zatwarnicki, K., Zatwarnicka, A.: Adaptive and Intelligent Request Distribution for Content Delivery Networks. Cybernetics and Systems 38(8), 837–857 (2007)
2. Borzemski, L., Zatwarnicki, K., Zatwarnicka, A.: Global Adaptive Request Distribution with Broker. In: Apolloni, B., Howlett, R.J., Jain, L. (eds.) KES 2007, Part II. LNCS (LNAI), vol. 4693, pp. 271–278. Springer, Heidelberg (2007)
3. Borzemski, L., Zatwarnicki, K.: Using Adaptive Fuzzy-Neural Control to Minimize Response Time in Cluster-Based Web Systems. In: Szczepaniak, P.S., Kacprzyk, J., Niewiadomski, A. (eds.) AWIC 2005. LNCS (LNAI), vol. 3528, pp. 63–68. Springer, Heidelberg (2005)
4. Borzemski, L., Zatwarnicki, K.: Fuzzy-Neural Web Switch Supporting Differentiated Service. In: Gabrys, B., Howlett, R.J., Jain, L.C. (eds.) KES 2006. LNCS (LNAI), vol. 4252, pp. 195–203. Springer, Heidelberg (2006)
5. Cardellini, V., Casalicchio, E., Colajanni, M., Mambelli, M.: Web Switch Support for Differentiated Services. ACM Perf. Eval. Rev. 29(2), 14–19 (2001)
6. Cardellini, V., Casalicchio, E., Colajanni, M., Yu, P.S.: The State of the Art in Locally Distributed Web-Server Systems. ACM Comp. Surv. 34(2), 263–311 (2002)
7. Casalicchio, E., Colajanni, M.: A Client-Aware Dispatching Algorithm for Web Clusters Providing Multiple Services. In: Proc. WWW10, pp. 535–544 (2001)
8. Lee, K.-M., Kwak, D.-H., Leekwang, H.: Tuning of fuzzy models by fuzzy neural networks. Fuzzy Sets and Systems 76(1), 47–61 (1995)
9. Mamdami, E.H.: Application of Fuzzy Logic to Approximate Reasoning Using Linguistic Synthesis. IEEE Trans. on Computers C-26(12), 1182–1191 (1977)
10. Markus, H., Beaumont, L.: Content Networking: Architecture, Protocols, and Practice. Morgan Kaufmann Publishers, USA (2005)
11. Menasce, D., Almeida, V.: Capacity Planning for Web Performance. Metrics, Models, and Methods. Prentice Hall, New York (1998)
12. Pai, V.S., Aron, M., Banga, G., Svendsen, M., Druschel, P., Zwaenpoel, W., Nahum, E.: Locality-Aware Request Distribution in Cluster-Based Network Servers. SIGOPS Oper. Syst. Rev. 32(5), 205–216 (1998)

# Digital Video Watermarking Based on RGB Color Channels and Principal Component Analysis

Hanane Mirza, Hien Thai, and Zensho Nakao

Department of Electrical and Electronics Engineering, University of the Ryukyus,
Okinawa 903-0213, Japan
{hanane,tdhien,nakao}@augusta.eee.u-ryukyu.ac.jp

**Abstract.** Embedding a digital watermark in an electronic document is proving to be a feasible solution for multimedia copyright protection and authentication purposes. In the present paper we propose a new digital video watermarking scheme based on Principal Component Analysis. We detect the video shots based on informational content, and color similarities; we extract the key frames of each shot and each key frame is composed of three color channels, and our proposed algorithm allows us to embed a watermark in the three color channels RGB of an input video file. The preliminary results show a high robustness against most common video attacks, especially frame dropping, cropping and recalling for a good perceptual quality.

**Keywords:** Multimedia protection, Video watermarking, PCA, Color channels.

## 1 Introduction

A picture is worth a thousand words. And yet, there are many phenomena which are not adequately captured by a single static photo. The obvious alterative to static photography is video. The video becomes an important tool for the entertainment and educational industry. However the entertainment industry is losing billions of dollars every year due to the new information marketplace where the digital data can be duplicated and re-distributed at virtually no cost. One possible solution to this problem is video watermarking. This involves the addition of an imperceptible and statistically undetectable signature to video file content. The embedded watermark should be resistant to common methods of signal processing, and, at the same time, it should not change the quality of the original video file.

Most of the proposed video watermarking schemes are based on the techniques of image watermarking. But video watermarking introduces some issues not present in image watermarking. Among the various video watermarking proposed schemes, Dittmann et al.[2] have embedded in the extracted feature of a video stream, while P.W Chan *et al. [1]* have used the Discrete Wavelet Transform by embedding in frequency coefficients of video frames. On the other hand

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 125–132, 2008.

Hien D Thai *et al [4]* were the first to introduce the PCA domain to gray-scale image watermarking.

In a previous work [5] we embedded the watermark in the three color channels of a color fixed image. In the present paper we tried to take advantage of the texture of video units to extract the key frames of the input video [6][7]; frames can be considered as color images. In this paper we propose to embed an imperceptible watermark separately, into the three different RGB channels of the video frame. We used the PCA transform to embed the watermark in each color channel of each frame. The main advantage of this new approach is that the same or multi-watermark can be embedded into the three color channels of the image in order to increase the robustness of the watermark. Furthermore, using PCA transform allows to choose the suitable significant components into which to embed the watermark.

## 2 Proposed Algorithm

### 2.1 Video Texture

Most of the existing effort has been devoted to the shot-based video analysis. However, in this work we will focus on the frame-based video analysis.

Video: An unstructured data stream, consisting of a sequence of video shots.
Scenes: Semantically related shots are merged in scenes.
Shots: Video units produced by one camera, and the shots boundary detection is made using the key frames. Shot boundary detection is important with respect to the trade-off between the accuracy and the speed in the reconstruction phase.
Frames: It is one complete scanned image from a series of video images; it is a static image. In the present paper we decompose the video[Fig.1] stream to sequences, then to scenes then to shots and then we extract each frame in each shot, using key frame extraction technique in [8] based on spatio-temporal features of the shots; we embed the watermark in each key-frame for robustness reasons.

### 2.2 Principal Component Analysis

In digital image processing field, the PCA or also called the KL transform, is considered as a linear transform technique to convey most information about the image to principal components. In the present algorithm, we first separate each frame to three color RGB channels, and we separately apply the PCA transform to each of the sub-frames before we proceed to the proper watermarking process. In fact we need to extract the principal component of sub pixels of each subframe by finding the PCA transformation matrix.

Each sub pixel is transformed by the PCA transformation matrix $[\varphi]$. It is then of primary importance to find the transformation matrix $[\varphi]$, going through the following process:

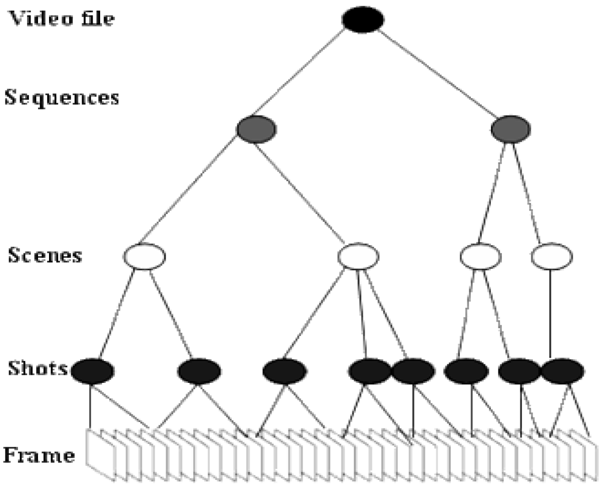


**Fig. 1.** A hierarchical video representation

*Task 1*: For numerical implementation and convenience we divide the frame F to a certain number of sub-frames. We consider each sub-frame an independent vector (vector of pixels). Thus, the frame data vector can be written as:$F = (f_1, f_2, f_3..., f_m))^T$where the vector $f_i$ is the $f^{th}$ sub image, $T$ denotes the transpose matrix, each sub-frame has $n^2$ pixels, and each vector $f_i$ has $n^2$ components.
*Task 2*: Calculate the covariance matrix $C_x$ of sub-frame, eigenvectors, and eigenvalues of the covariance matrix.

$$C_x = E(F - m_i)(F - m_i)^T \tag{1}$$

where $m_i = E(F)$ are the mean vector of each sub-vector $f_i$ , each sub-picture may now be transformed into uncorrelated coefficients by first finding the eigenvectors (basic functions of transformation) and the corresponding eigenvalues of the covariance matrix:

$$C_x\Phi = \lambda_x\Phi \tag{2}$$

The basis function $[\varphi]$ is formed by the eigenvectors $\Phi = (e_1, e_2, e_3...e_{n^2})$ . Eigenvalues $\lambda(\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq ..... \geq \lambda_{n^2})$ and eignvector $[\varphi]$ are sorted in descending order. The matrix $[\varphi]$ is an orthogonal matrix called basis function of PCA.
*Task 3*: Transform sub-frame into PCA component. The PCA transform of sub-frame can be done by the inner product of the sub-frame with the basis functions. The original frame F can be de-correlated by the basis function frame $[\varphi]$, and we obtain Y by the following equation:

$$Y = \Phi^T F = (y_1, y_2, ...y_m)^T$$

The corresponding values are the principal components of each sub-frame. Corresponding to each sub-frame, we can embed the watermark into selected components of sub-frame.
*Task 4*: To retrieve the watermarked frame, we perform the inverse process using the following formula:

$$F = (\Phi^T)^{-1} Y = \Phi Y \tag{3}$$

### 2.3 Embedding Process

In this work our encoding process consists of the following steps:

***First step:*** An input video is split into audio and video stream [Fig.2] and the video stream is represented by the key-frame [Fig1]. Each frame is considered as a color image separately.
***Second step:*** In order to embed a watermark into a given original color frame of size F(N, N), using the proposed technique, we have to separate the frame F(N, N) to three RGB color channels: Red, Green and Blue. We get, respectively, the three sub-frames:$F_R(N, N)$, $F_G(N, N)$ and $F_B(N, N)$.[Fig 3]
***Third step:*** For each of the three sub-frames we apply PCA transform. Each of the three color-banded frames $F_R$ , $F_G$ and $F_B$ is separately subdivided to a certain number n of sub-frames[Fig.3]. We can get PCA basis function for each of the sub-frames respectively: $[\Phi]_R$, $[\Phi]_G$, and $[\Phi]_B$. The principal components of each of $F_R$, $F_G$ and $F_B$ are computed by the process discussed above through task 1 to task 3. We then have the three PCA coefficients : $Y_R$, $Y_G$, $Y_B$.
***Fourth step:*** Select the perceptually significant components of each of the three coefficients, into which the watermark will be inserted. In this algorithm, the watermark is a random signal that consists of a pseudo-random sequence of length M, the values of w is a random real number with a normal distribution, $W = w_1, w_2 ... w_M$. We need then to embed the watermark into the predefined components of each PCA sub-block uncorrelated coefficients. The embedded coefficients were modified, for each sub-frame, by the following equation:

$$(y_i)_w = y_i + \alpha \mid y_i \mid w_i \tag{4}$$

where $\alpha$ is a strength parameter. Then we obtain $y_{wR}$, $y_{wG}$, $y_{wB}$.
***Fifth step:*** The three RGB watermarked color channels are separately recovered by the inverse PCA process. (Task 4.)

$$F_w = (\Phi^T)^{-1} Y_w = \Phi Y_w \tag{5}$$

And by superposing the three resulting color channels $F_{wR}$, $F_{wG}$ and $F_{wB}$ we retrieve the watermarked frame $F_w(N, N)$.
***Sixth step:*** We proceed to video reconstruction, by retrieving first the video shots [7], we reintegrate the watermarked key frames in the order they originally were, and by using the Video/ audio merger tool, we reproduce the watermarked video file.

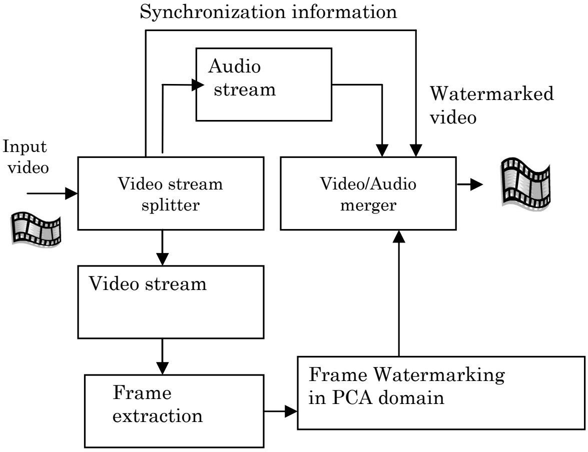


**Fig. 2.** Video watermarking algorithm

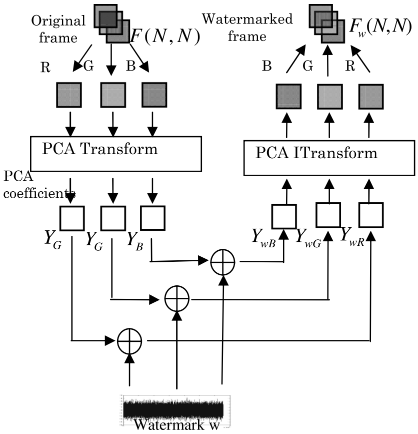


**Fig. 3.** Key Frame watermarking process in PCA domain

### 2.4 Decoding Process

For recognition of the authenticity of the embedded watermark, a watermark is detected through the process described in [Fig.4]. The tested video stream is subjected to frames extraction process [Fig.1], and for each frame we applied the correlation based detection. Three extracted watermarks are compared to other 1000 watermarks. Suppose we received an image, and we need to confirm the positive or negative presence of the original watermark in the watermarked image $F^*(N, N)$. For $F^*(N, N)$ we apply step 1 and step 2 (as detailed in the encoding process). In consequence we get the PCA coefficient for each of $F^*_R(N, N)$,

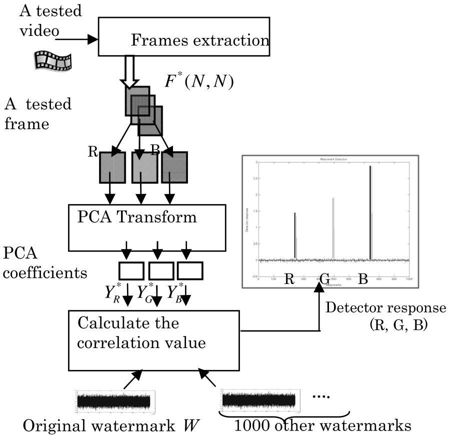


**Fig. 4.** Watermark detection process

$F_G^*(N,N)$, $F_B^*(N,N)$, namely; $Y_R^*$, $Y_G^*$, $Y_B^*$. The correlation formula used, for each sub-frame separately is:

$$(CV) = \frac{WY^*}{M} = \frac{1}{M}\sum_{i=1}^{M} w_i y_i^* \tag{6}$$

## 3 Computer Simulation

For an MPEG video of 15 minutes extract of the movie "rush hour 2", of rate 30 frame/ second, and resolution 640x480, we extract 98 color key frames. We randomly generate an M=65536 length watermark . After extracting all the 98 color frames, we proceed to watermarking process as described in sub-section (2.3) with strength parameter $\alpha = 0.7$, the watermarked frames were uploaded to a video editor (Honestec Video Editor) for reintegration of the key frames, Figure.5 shows an original and watermarked frame number 19 as examples, more results to come in the complete paperversion.

After applying the proposed watermark to the video stream, the obtained watermarked and reconstructed video shows that there is no noticeable difference between the watermarked and the original video, which confirm the invisibility requirement in our watermarking method. (An average PSNR value is shown in Table 1). In order to test the the robustness of our algorithm, a number of signal processing attacks were applied to the watermarked video stream as described

**Table 1.** Average PSNR and detection rate for watermarked frames

| | |
|---|---|
| Detection rate | $R_w = 0.67, G_w = 0.70,$ $B_w = 0.80$ |
| Frame PSNR(average) | 83.2dB |

**Fig. 5.** Original(left) and watermarked (right) key frame $N^o19$

in section 2.4, and the system shows good results for watermark detection. From Table 2 we can see that for cropping, frame dropping and rotation attacks we could easily detect the presence of the three watermarks in the three color layers, and an overall watermark was calculated for comparison. As for both median filtering and rescaling attacks, at least one of the three watermarks was detected which demonstrates the effectiveness of the system. The overall watermark detection after attacks, using StirMark, are shown in Table 2 along with a comparison with the video watermarking schemes previously proposed, where:

(a)The proposed method: Color channels video watermarking based on PCA
(b)DWT- based watermarking scheme[1].
(c)Scene-based watermarking scheme.
(d)Visual-audio hybrid approach.

**Table 2.** Attacks and comparison with previously developed schemes

| Attack/class | a | b | c | d |
|---|---|---|---|---|
| PSNR | 83.2 | 72.0 | 76.0 | 83.0 |
| Cropping | 0.73 | 0.68 | 0.66 | 0.78 |
| Rescaling | 0.65 | 0.63 | 0.62 | 0.75 |
| Frame dropping | 0.91 | - | - | - |
| Rotation | 0.71 | 0.60 | 0.61 | 0.73 |
| Median Filter | 0.63 | 0.54 | 0.54 | 0.74 |

## 4 Conclusions

A new digital video watermarking technique is proposed in this paper. The idea of embedding the watermark in the three color channels of each key frame, was checked for robustness by inserting it in each color channel while the PCA based watermarking scheme allowed to select the appropriate PCA coefficients for embedding, and in fact we could demonstrate that it is always possible to watermark a color video file without affecting its perceptual quality.

## Acknowledgments

This research was supported in part by Ministry of Internal Affairs and Communications (Japan) under Grant: SCOPE 072311002, for which the authors are grateful.

## References

1. Chan, P.W., Lyu, M.R.: A DWT-based Digital Video Watermarking Scheme with Error Correcting Code. In: Proceedings Fifth International Conference on Information and Communications Security, pp. 202–213 (2003)
2. Dittmann, Steinebach: Joint Watermarking of Audio-Visual Processing. In: IEEE Fourth Workshop on Multimedia Signal Processing, France (October 2001)
3. Bloom, J.A., Cox, I.J.: Copy Protection for DVD Video. Proceedings of the IEEE, USA (1999)
4. Hien, T.D., Chen, Y.-W., Nakao, Z.: A robust digital watermarking technique based on principal component analysis. International Journal of Computational Intelligence and Applications 4(2), 138–192 (2004)
5. Miyara, K., Hien, T.D., Harrak, H., Nakao, Z., Nagata, Y.: Multichannel color image watermarking using PCA eigenimages. In: Advances in soft computing, vol. 5, pp. 287–296. Springer, Heidelberg (2006)
6. dl Sch, A., Szeliski, R., Salesin, D.H., Essa, I.: Video textures. In: Proceedings of SIGGRAPH 2000, pp. 489–498 (2000)
7. Rui, Y., Huang, T., Mehrota, S.: Exploring Video structure beyond the shots. In: Proceedings of IEEE International Conference on Multimedia Computing and Systems, USA (1998)
8. Bruno, J., Bruno, E., Pun, T.: Information-theoritic temporal segmentation of video and applications:multiscale keyframe selection and shot boundaries detection. Kluwer Academic Publishers, Netherland (2005)

# A Watermarking Approach for MIDI File Based on Velocity and Duration Modulation

Alexander Adli[1], Hanane Mirza[1], and Zensho Nakao[2]

[1] Graduate School of Engineering & Science, University of the Ryukyus,
Okinawa 903-0213, Japan
[2] Department of Electrical & Electronics Engineering, University of the Ryukyus,
Okinawa 903-0213, Japan
{alex, hanane, nakao}@augusta.eee.u-ryukyu.ac.jp

**Abstract.** The feasibility of inserting a signature into a musical file is proven to be one of widely used technique for copyright and security purposes. In this paper we present a brief overview of the watermarking approaches and implementation in simple MIDI files (SMF). The authors suggest a method based on velocity and duration modulation to embed a digital logo representing watermark data. The watermark robustness is shown via experiments with simulated attacks. Potential problems are discussed in further studies.

**Keywords:** Watermark, Symbolic domain, velocity modulation, duration modulation, MIDI parameters.

## 1 Introduction

MIDI (Musical Interface Digital Instrument) files are one of the most popular musical performance files. They are often used by composers and musicians. These files are often used on the end-user level where the user might use them simply for music listening or for further musical processing. Musicians who use digital instruments (e.g. digital piano) for recording might record and distribute their music as MIDI files.

The information about the author of the MIDI files is written in special tags inside the file. However this information can be easily deleted of modified by a third party. A watermark embedded within the music information seems to be an adequate solution. Until today there is no wide used watermark application or technique applicable on MIDI files to protect their ownership.

This paper provides a short survey about applications and methods which handled this problem. The authors suggest a novel technique for watermark in MIDI using two algorithms – velocity modulation and duration modulation. Although the watermark size is relatively small but it has been proved to have higher robustness than the previously existed methods.

A simulation is provided to support the methods, and a few problems are mentioned in the conclusions and further studies.

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 133–140, 2008.

## 2 Watermarking Techniques in MIDI

MIDI files are music containers for the musical performance. Unlike the popular digital music representation formats where the wave is described (both compressed and uncompressed) the MIDI format is simply commands for the synthesizer and it does not contain any information about the sound.

As a result the MIDI file can be viewed somehow like text file, which makes the watermarking approaches of image files and digital audio files not directly applicable. The structure of the file has to be considered to avoid corruption of the content.

Basically saying the MIDI file is a stream of note commands. Every command consists of time delay, command code, channel number, note number and velocity. When the note is to be released usually the same command is used with a velocity equals *0*. Along with these commands there are meta commands for non musical information and special system commands which are not standardized.

Several techniques for watermark embedding into the MIDI files exist. Jana Dittmann describes a framework for watermarking in MIDI for eCommerce application[1]. In [2] an algorithm using the Least significant bit (LSB) technique applied to the velocity parameters was used for steganography purposes. A velocity number is given with every note transmitted via the MIDI system. It describes, how fast the key on the keyboard belonging to the note has been pressed. The faster one presses the key, the louder the note will be played.

Another technique is based on using the redundancy in MIDI files, specifically, when two consecutive commands notes shares the same command code and the same channel, which is the general case, then the byte describing the channel number and command code is usually but not necessarily omitted to save space. MIDI players still recognizes the commands correctly. This feature of MIDI was used in [2] for steganography purposes, and can be applied for watermark. However, even unintentional attack of simply resaving the file using any sequencer might lead to a total loss of this information.

Adding virtual notes is a watermark technique in MIDI, and is applied by Changsheng Xu in [3].The watermark is embedded into articulation parameters of the generated virtual MIDI notes. Virtual MIDI notes mean that these notes will not affect the original MIDI quality.

Although the virtual notes are robust watermark as they cannot be deleted by any applied filter on the MIDI file, the musicians who use the MIDI files consider the virtual notes totally unacceptable because they affect the file looks like when it is viewed as piano roll or musical sheet which is the most common way of viewing these files. In addition to that adding virtual notes affects the MIDI oriented music analysis automated systems and the algorithms they uses like in [4].

The watermark information can be inserted by adding additional meta commands to the MIDI stream. The meta commands provides additional (non musical) information about the musical piece. The meta commands contain information about the author. Meta commands can be intentionally deleted and the watermark hidden using them will be lost.

Yamaha Corporation described a product called MIDStamp [5] where it claimed that it can embed a watermark into the MIDI file and which is robust toward attacks, however this product was discontinued before its release.

# 3 Our Watermark Embedding Methods

Our watermark embedding technique in MIDI files is based on time modulating and velocity variations. We avoided using virtual notes which lead into "visual" corruption of the file when presented as musical sheet, we also avoided using meta or system commands which can be easily deleted and makes the watermark less robust.

This watermark approach is more robust when compared with the algorithms which use the LSB of the velocity parameter or the time parameter. It does not apply any visual changes on the MIDI files if represented as musical sheet and the file does not lose its musical quality.

This is not blind watermarking, and the original file is required to restore the watermark.

Below we describe these two approaches in details and simulate attacks.

## 3.1 Note Duration Modulating Approach

Every command has delta time information describing who long this command is to be executed later than the previous one. However, a better representation might be Onset time with the duration as is usually used to show MIDI files as list of notes.

The ear is too sensitive to Onset time changing, that even small modifications can be sensed by musicians. The duration is used in our method instead of onset to hide the watermark information.

The algorithm basically uses two consecutive notes' duration $D_{1original}$, $D_{2original}$ values to hide *1* bit of information. These two values are modified in opposite directions to get $D_{1water}$, $D_{2water}$. Formula (1) writes the bit value "*1*" and (2) writes the bit value "*0*" as follows:

$$\begin{cases} D_{1water} = D_{1original} - c \\ D_{2water} = D_{2original} + c \end{cases} \tag{1}$$

$$\begin{cases} D_{1water} = D_{1original} + c \\ D_{2water} = D_{2original} - c \end{cases} \tag{2}$$

where *c* is a small value which depends on the tempo.

To restore the watermark, the watermarked file is compared with the original file and if $\frac{D_{1water}}{D_{1original}} > \frac{D_{2water}}{D_{2original}}$ then the value "1" is extracted. Otherwise "0" is extracted as $\frac{D_{1water}}{D_{1original}} \le \frac{D_{2water}}{D_{2original}}$ is true.

The proportions comparison instead of merely comparing the values between the original file and the watermarked file provides better watermark robustness against tempo changing attacks and other attacks.

To avoid the position of the watermark from been easily calculable by the attacker, a simple technique is implemented. A note's duration is considered as a candidate for the first watermark bit embedment if and only if its lowest decimal digit value is odd. The next watermark bit is another couple of note's duration where the lowest decimal digit is even, the next is odd again and so on. This is done to keep the proportion of odd and even values unchanged in the watermarked file. As the value is changed due to the watermark, it is impossible to know the original position of the watermark without comparing with the original file which is not distributed. See Table 1.

**Table 1.** As the first note's duration is even it is not used for the watermark purpose. The next note (63) has an odd value of its duration, and thus is used along the next one for the watermark purpose, and they are modified according to (1).

| OnSet Time | Note Number | Velocity | Original duration | Watermark duration |
|---|---|---|---|---|
| 1,05 | 60 | 43 | 0,82 | 0,82 |
| 1,26 | 63 | 47 | 0,77 | 0,76 |
| 2,16 | 65 | 44 | 0,61 | 0,62 |

In this way this technique is relatively robust for time quantization attacks and to tempo change attacks. Any information written using LSB algorithm would have been lost after quantization.

A time quantization attack usually alters the Onset and duration values in a small and random value. In case of linearly distributed random quantization of a value *c,* duration is increased or reduced by value *c* with probability $p=\frac{1}{2}$. Such quantization will corrupt the watermark bit with a probability $p=\frac{1}{4}$ (two values have to be modified in specific direction).

An attack like crescendo or diminuendo or tempo changing does not noticeably harm the watermark as the duration proportions keeps same.

To minimize the lost of information due to quantization attack an error correcting code can be used. We tried the using of the very simple forward error correction method (FEC) which is the triple bit demarcating vote. The value is considered "0" if number of zeros in the triple is more the ones and vice versa. This method is considered acceptable as it is designed for cases where error positions are unknown and random.

Regarding the watermark storage; as an average, three notes store 1 bit, and three bits codes one watermark bit (via FEC). An average piece of 1000 notes can encode only about 112 bits or 14 bytes of watermark information.

### 3.2 Velocity Modulation Approach

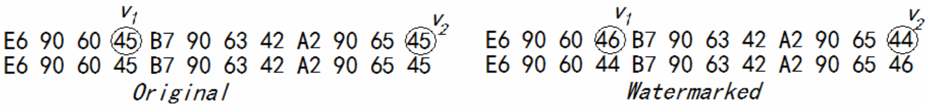
Every note on command has a 7 bit velocity parameter. Velocity describes how "loud" the specific musical note is.

The algorithm does the following: for any note which has velocity within the working range (2 to 126), it finds the closest latter note to it which has the same velocity value. The two notes are now used together to encode 1 bit. Formula (3) is used to encode the binary value "1" and formula (4) is used to encode value "0".

$$\begin{cases} v_{1water} = v_{1original} - c \\ v_{2water} = v_{2original} + c \end{cases} \tag{3}$$

$$\begin{cases} v_{1water} = v_{1original} + c \\ v_{2water} = v_{2original} - c \end{cases} \tag{4}$$

where $v_{\text{1original}}$,$v_{\text{2original}}$ are the original and equal velocities of the two notes, $v_{\text{1water}}$,$v_{\text{2water}}$ are the watermarked values, and $c$ is a value which depends on velocities dispersion, $c \geq 1$. Figure 1 illustrates 1 bit of watermark embedded to the file by changing two velocity parameters.



```
E6 90 60 45 B7 90 63 42 A2 90 65 45    E6 90 60 46 B7 90 63 42 A2 90 65 44
E6 90 60 45 B7 90 63 42 A2 90 65 45    E6 90 60 44 B7 90 63 42 A2 90 65 46
```

**Fig. 1.** The modified velocities in the watermarked file encodes "0" in the first line and "1" in the second line

The position of the watermark is unknown for the attacker as the originally equal velocities are now modified.

To restore the watermark, the file is compared with the original file, and a bit "0" is revealed if $v_{\text{1original}}$= $v_{\text{2original}}$ in the original file corresponds to $v_{\text{1water}} > v_{\text{2water}}$ in the watermarked file, and "1" if $v_{\text{1water}} < v_{\text{2water}}$.

This provides much more robust watermark especially when compared with the LSB method usually applied on the velocities to hide the data.

A common attack on velocity is decreasing or increasing all values, or changing the contrast by changing the dispersion. In all these attacks the LSB information would have been totally lost where our methods perverse the watermark untouched.

Another possible attack is the quantization of velocities. Usually this means randomly adding or subtracting small values to the velocity parameter. Sometimes this is used to add "humanity" to the automated performance.

If the original file were watermarked using *c=1*, and a quantization algorithm randomly adds or subscribes a value $c$=1 from the velocity parameters $v_{\text{1water}}$, $v_{\text{2water}}$, of the watermarked file, then such a case where this quantization reveres the original reading can take place only if $v_{\text{1water}}$, $v_{\text{2water}}$ are modified in different direction (their distance is 2$c$ = 2) which has the probability $p = \frac{1}{4}$. In all the other cases the sign between $v_{\text{1water}}$ and $v_{\text{2water}}$ will not be changed.

The lost information can be restored if an error-correcting code is used.

Regarding the watermark storage, a file with 1000 notes has in average 90 percent of its velocities within the range 40 and 120 i.e. 900 notes. Out of these notes we will expect around 40 notes (80/2) cannot be grouped in couples of equal velocities. As one couple is used to embed 1 bit, the total storage is about 430 bits. If FEC triple bit is used then it is around 144 bit or 18 byte which is still more than previous method of duration modulation.

## 4 Simulations and Results

These methods were realized in Matlab using MIDITools[6]. Several functions were developed in C++ using the MIDI classes[7].

The two methods are used together to embed the watermark information. The watermark is chosen to be a binary digital (ones and zeros) logo with size *28x28* or 784 bits (see fig.3 A).

The original file is Grieg's wedding day in Trollhaugen played on a digital Yamaha piano and recorded via MIDI controller as a MIDI file. The file contains around 3600 notes (around 7 minutes) which mean it is suitable to fit around 918 watermark bits (around 400 bits as a duration modulation and around 518 bits in velocities) which is pretty enough to store our 784 bit watermark.

The watermark is then embedded into the original file (fig. 2). According to expert conclusion and our expectation there are audible differences between the original and watermarked files.

The watermarked file was then modified: meta events about author were deleted, the time constant were modified (this causes all the time stamps to be changed), the tempo had been made a little faster, all the notes were *1* semitone up transposed, a quantization were applied on all velocities, onsets and durations, all notes' channel numbers were changed. As a result, the attacked file has almost all his bytes modified when compared with the watermarked file.

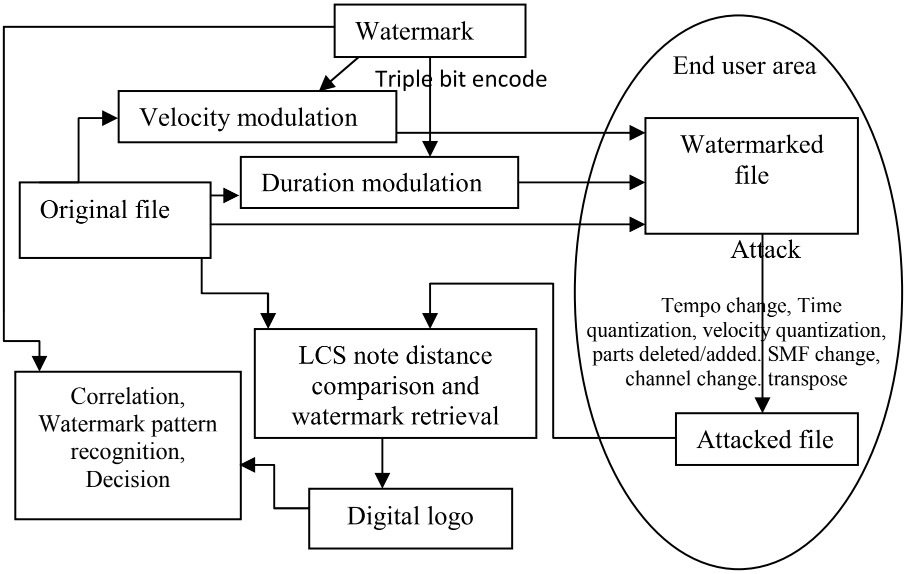

Watermark
Triple bit encode
End user area
Velocity modulation
Watermarked file
Duration modulation
Original file
Attack
Tempo change, Time quantization, velocity quantization, parts deleted/added. SMF change, channel change. transpose
LCS note distance comparison and watermark retrieval
Correlation, Watermark pattern recognition, Decision
Attacked file
Digital logo

**Fig. 2.** General watermark embedding and detection scheme after attacks

In the restoration step, the attacked file is compared with the original file (non distributed file) using the longest common subsequence (LCS) method [8] applied on note distances (see table 2).

**Table 2.** The LCS implementation. The distance is preserved between the original and attacked files although the latter were one semitone transposed. The LCS also did not include the deleted not in the subsequence to compare.

| | Original string | Attacked string (transposed, 1 deleted) |
|---|---|---|
| Notes | *60 63 65 72 75 77 84 87 89* | *61 64 66 73 76 78 88 90* |
| Distances | *3 2 7 3 2 7 3 2* | *3 2 7 3 2 10 2* |
| LCS | *3 2 7 3 2 2* | |

The LCS applied on distances allows the two file to be compared even if notes are edited or parts are cut or transposed. The relative note distances cannot be changed as this completely "ruins" the music.

The longest common subsequence of the attacked file and the original file are compared. The watermark is restored after all the attacks. (Fig 3.)

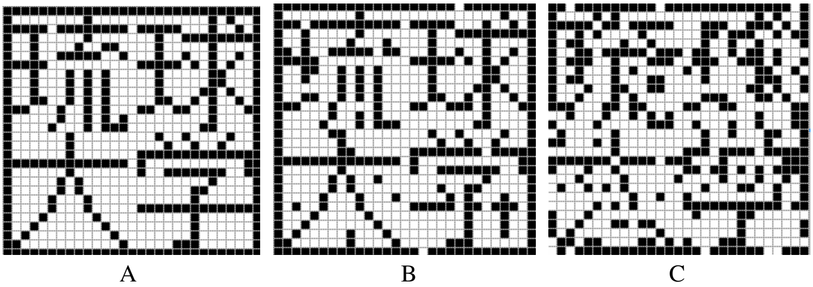


**Fig. 3.** The watermark. A(left) – the original watermark (University of the Ryukyus written in Japanese), B(middle) – the restored watermark from the attacked file with the FEC triple bit error correcting method used, C (right) – the restored watermark without any error correction.

The watermark which was embedded using both methods (duration time modulating and equal-velocity modulating) with a triple bit error correction method is restored as shown in "B" and any Japanese who can read the original is very likely to be able to read "B". Although "C" (without error correction) is almost impossible to read, it still strongly suggests the existence of the watermark.

## 5 Conclusions and Further Work

This paper provides a description on watermark techniques in the popular symbolic music files – MIDI format. The authors describe their technique to embed a watermark into the MIDI file using equal velocity modulation and note duration modulation. The suggested technique does not degrade the musical quality of the file, it does not affect the visual representation of the file, and does not include information which might negatively affects automated music analysis. No extra commands are added, using meta commands and system commands is avoided. The watermark position cannot be detected by the attacker, and it is robust for Sequencers attacks including, resaving the file, SMF format changing, transposing, channel change, velocity quantization, time quantization, tempo changing, adding crescendos and diminuendos, adding or deleting single notes.

The average watermark capacity is around 1 bit per 4 notes.

The robustness is partially achieved by applying forward error correcting method of triple bit error correcting code.

The watermark restoration requires the original file, and the two file are compared using longest common subsequence method.

The experiments show that the watermark is readable after multiple attacks when FEC is used, and detectable when FEC is not used.

However, the watermark is not totally robust, and it can be vulnerable for some attacks with intension of destroying it. For example a modification of any consecutive durations in opposite directions might eventually destroy most of the watermark. To avoid this, the couple might be scattered in the file instead of being consecutive (make it like the scattered velocity values).

Another way is to destroy the ability of restoration algorithm to restore the watermark. For example adding virtual notes between any 2 notes will make the LCS unable to function properly. Probably a modified version of LCS and optimized for MIDI files can introduce a solution.

**Acknowledgments.** This research was supported in part by Ministry of Internal Affairs and Communications (Japan) under Grant: SCOPE 072311002, for which the authors are grateful.

## References

1. Dittmann, J., Steinebach, M.: A framework for secure MIDI eCommerce, German National Research Center for Information Technology, Darmstadt, Germany (2002)
2. Adli, A., Nakao, Z.: Three Steganography algorithms for MIDI files. In: Proceedings of 2005 International Conference on Machine Learning and Cybernetics, Guangzhou, China, August 2005, vol. 4, pp. 2404–2407 (2005)
3. Xu, C., Zhu, Y., Feng, D.D.: Content Protection and Usage Control for Digital Music. In: Proceedings of the First International Conference on WEB Delivering of Music, USA (2001)
4. Ozcan, G., Isikhan, C., Alpkocak, A.: Melody Extraction on MIDI music Files. In: Proceedings of the seventh IEEE International Symposium on Multimedia (2005)
5. Yamaha corporation, MIDStamp (1998), `http://www.yamaha.co.jp/english/news/98090302.html`
6. Eerola, T., Toiviainen, P.: MIR in Matlab. In: Proceedings of International Symposium in Music Information Retrieval (2004)
7. Adli, A., Nakao, Z.: MIDI Classes for Symbolic Music Processing (under review)
8. Hakata, K., Imai, H.: The longest common subsequence problem for small alphabet size between many strings. In: Ibaraki, T., Iwama, K., Yamashita, M., Inagaki, Y., Nishizeki, T. (eds.) ISAAC 1992. LNCS, vol. 650. Springer, Heidelberg (1992)
9. MIDI Manufacturers Association, `http://www.midi.org`
10. Robert Rowe: Machine Musicianship. The MIT Press, Cambridge (2001)

# Application of Interactive Genetic Algorithms to Boid Model Based Artificial Fish Schools

Yen-Wei Chen[1,2], Kanami Kobayashi[2], Hitoshi Kawabayashi[2], and Xinyin Huang[3]

[1] Electronics & Information Eng., School, Central South Univ. of Forestry and Tech., Changsha, 410004, China
[2] College of Information Science and Engineering, Ritsumeikan University, Japan
[3] School of Education, Soochow University, Suzhou, Jiangsu 215006, China
chen@is.ritsumei.ac.jp

**Abstract.** In this paper, we present an extended boid model for simulating the aggregate moving of fish schools in a complex environment. The boids model is an example of an individual-based model. The global behavior of the school is simulated by a large number of interesting individual boid (fish). In our proposed model, each boid is an agent that following six behavior rules: avoiding collision against schoolmates; gathering together; following a feed; avoiding obstacle; avoiding enemy boids; boundaries. The moving vector of each boid is a linear combination of five behavior rule vectors, and the coefficients are optimized by using an interactive genetic algorithm (IGA). Unlike the classical GA, interactive GA can adopt a user's subjective evaluation as fitness, which is useful when a fitness function cannot be exactly determined.

**Keywords:** Boid model, interactive genetic algorithm (IGA), fish school, aggregate moving, behavior rule.

## 1 Introduction

Simulating the aggregate moving of a fish school or a bird flock is an important issue in the areas of computer animation and artificial life. In 1986, Relond proposed a computer model of coordinated animal motion such as bird flocks and fish schools, which is called as boids [1]. The Boids model has three basic behavior rules, which are avoiding collision against neighbors; matching and coordinating own moves with neighbors; gathering together. The boids model has been used for modeling of fish [2]. In this paper, we present an extended boids model for simulating the aggregate moving of fish schools in a complex environment. Three behavior rules are added to the extended boids model: following a feed; avoiding obstacle; avoiding enemy boid. Each rule is represented by a vector. The direction and amplitude of the vector are adaptive to the environment. The moving vector of the boid (fish) is a linear combination of every behavior rule vector. As increasing the behavior rules, the setting of the coefficients becomes complex and difficult. In our previous work [3], we proposed to

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 141–148, 2008.

apply a genetic algorithm (GA) [3,4] to optimize the coefficients. The drawback of the classical GA is that it is difficult to define a mathematic function which can be used as a measure of fitness. In this paper, we propose a new approach based on interactive GA (IGA) for optimization of boid models. Unlike the classical GA, interactive GA can adopt a user's subjective evaluation as fitness, which is useful when the fitness function cannot be exactly determined. Experimental results show that by using the IGA-based optimization, the aggregate motions of fish schools become more realistic and similar to behaviors of real fish world.

The paper is organized as following: the extended boid model is presented in Sec.2, the interactive genetic algorithm (IGA) for optimization of coefficients is presented in Sec.3 and the experimental results are shown in Sec.4. Finally, the conclusion is given in Sec.5.

## 2 Extended Boid Model

The boids model is an example of an individual-based model. Each simulated boid (fish) is implemented as an independent actor that navigates according to its local perception of the dynamic environment. The global behavior of the school is simulated by a large number of interacting individual boid (fish). In the extended boids model, each boid is an agent that follows following five behavior rules: avoiding collision against schoolmates; gathering together; following a feed; avoiding obstacle; avoiding enemy boids. The first two rules are Reynold's and the last three rules are our newly proposed ones.

### 2.1 Avoiding Collision against Schoolmates

The first rule is avoiding collision against schoolmates. The rule is illustrated in Fig.1. The vector determined by the first rule is shown in Eq.(1).

$$\mathbf{V}_1 = \begin{cases} \left(\dfrac{|\mathbf{BoidVec}|}{fKeepDist} - 1\right) \cdot \dfrac{\mathbf{BoidVec}}{|\mathbf{BoidVec}|}, & \left(|\mathbf{BoidVec}| \le fVisibleDist\right) \\ 0, & \left(|\mathbf{BoidVec}| > fVisibleDist\right) \end{cases}, \qquad (1)$$

where *fVisibleDist* is the visible distance of the boid (fish), *fKeepDist* is the safe distance for avoiding collision against schoolmates, and **BoidVec** is the vector from the boid to the nearest schoolmate. As shown in Eq.(1), when the distance to the nearest schoolmate is smaller than *fKeepDist*, a vector (force) is acted in opposite direction in order to keep away from the schoolmate. On the other hand, when the distance to the nearest schoolmate is larger than *fKeepDist*, a vector (force) is acted in the same direction in order to close to the schoolmate.

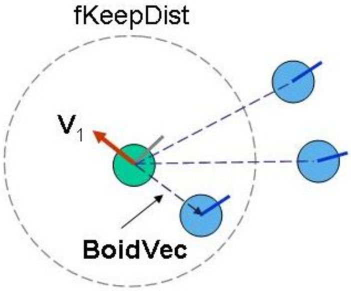


**Fig. 1.** Rule 1: avoiding collision against schoolmates

## 2.2 Gathering Together

The second rule is gathering together. A vector (force) is acted in the direction to the center (average position) of the neighborhood (fish school) in the view as shown in Fig.2. The vector is given by Eq.(2).

$$:\begin{cases} \dfrac{\mathbf{CenterVec}}{|\mathbf{CenterVec}|}, & \left(|\mathbf{CenterVec}| \le fVisibleDist\right) \\ 0, & \left(|\mathbf{CenterVec}| > fVisibleDist\right) \end{cases}, \qquad (2)$$

where **CenterVec** is the vector from the boid to the center of the neighborhood.

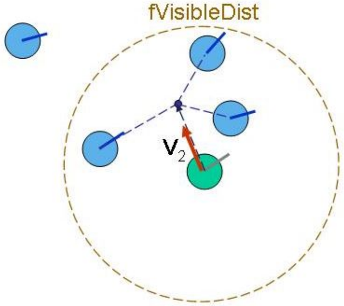


**Fig. 2.** Rule 2: gathering together

## 2.3 Following a Feed

The third rule is following a feed. A vector (force) is acted in the direction to the feed as shown in Fig.3. The vector is given by Eq.(3).

$$\mathbf{V}_3 = \frac{\mathbf{FoodVec}}{|\mathbf{FoodVec}|}, \qquad (3)$$

where **FoodVec** is the vector from the boid to the feed.

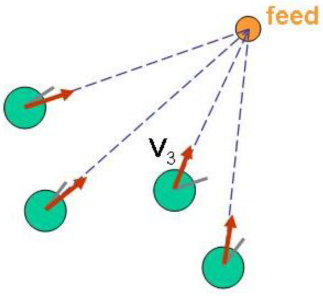


**Fig. 3.** Rule3: following a feed

## 2.4 Avoiding Obstacles

The fourth rule is avoiding obstacles. Obstacle avoidance allowed the boids to fly through simulated environments while dodging static objects. The rule is illustrated in Fig.4. Assuming the avoiding angle be α and the size of obstacle be *ObsMag*.

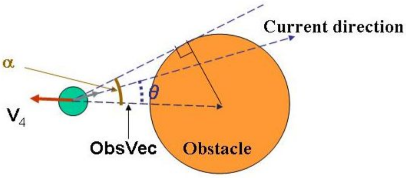


**Fig. 4.** Rule4: avoiding obstacles

$$\cos\alpha = \frac{\sqrt{|\mathbf{ObsVec}|^2 - \left(\dfrac{ObsMag}{2}\right)^2}}{|\mathbf{ObsVec}|}, \qquad (4a)$$

where **ObsVec** is the vector from the boid to the center of obstacle as shown in Fig.4. The vector acted for avoiding obstacle is given as

$$\mathbf{V}_4 = \begin{cases} -\cos\theta \cdot \left(1 - \dfrac{|\mathbf{ObsVec}|}{fVisibleDist}\right) \cdot \dfrac{\mathbf{ObsVec}}{|\mathbf{ObsVec}|} & \left(\cos\theta \ge \cos\alpha\right) \\ 0 \quad \left(\cos\theta < \cos\alpha\right) \end{cases}, \qquad (4b)$$

where θ is the angle of current direction of the boid with the obstacle.

### 2.5 Avoiding Enemy Boids

The fifth rule is avoiding enemy boid. When the boid finds an enemy boid in the visible distance, a vector (force) is acted in the opposite direction to the enemy boid as shown in Fig.5. The vector is given by

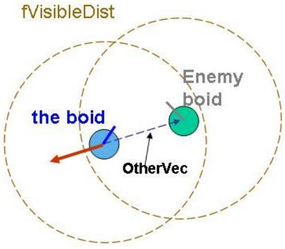


**Fig. 5.** Rule 5: avoiding enemy boids

$$\mathbf{V}_5 = \begin{cases} \left(\dfrac{|\mathbf{OtherVec}|}{fVisibleDist} - 1\right) \cdot \dfrac{\mathbf{OtherVec}}{|\mathbf{OtherVec}|}, & \left(|\mathbf{OtherVec}| \le fVisibleDist\right) \\ 0, & \left(|\mathbf{OtherVec}| > fVisibleDist\right) \end{cases}, \tag{5}$$

where **OtherVec** is the vector from the boid to the enemy boid.

### 2.6 The Moving Vector

The moving vector of each boid is determined by above five rules. The moving vector can be considered as a linear combination of the five vectors as

$$\mathbf{V} = w_1\mathbf{V}_1 + w_2\mathbf{V}_2 + w_3\mathbf{V}_3 + w_4\mathbf{V}_4 + w_5\mathbf{V}_5, \tag{6}$$

where $w_i$ is the coefficients used to balance the five rules and the coefficients should be optimized.

## 3 Interactive Genetic Algorithm (IGA)

As shown in Eq.(6), the moving vector of each boid is a linear combination of five vectors which are determined by each rule. As increasing the behavior rules, the setting of the coefficients becomes complex and difficult and they should be optimized.

Genetic algorithm (GA) is an optimization technique, which applies the principles of evolution found in nature to the problem of finding an optimal solution [3, 4]. Compared with conventional gradient-based optimization methods, since GA starts with a population of candidate solutions (multi-point search), it is easy to find a global optimum as shown in Fig.6. The GA has been applied in many fields including image processing [5-7]. In our previous work [3], we have proposed to apply the GA to optimize the coefficients. Since the aim of the optimization is to make the aggregate motions of fish schools more natural and more realistic, it is difficult for the classical GA to define a suitable mathematic function which can be used as a measure for evaluation or fitness. In this paper, we propose a new approach based on interactive GA (IGA) for optimization of boid models. Unlike the classical GA, interactive GA

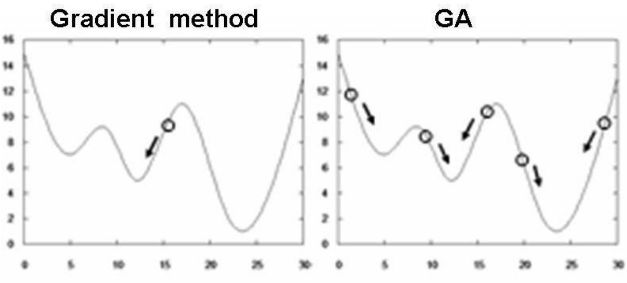


**Fig. 6.** Comparison of the GA with the gradient method

can adopt a user's subjective evaluation as fitness. Users evaluate the distance between the system output and the gaol in a psychological space. It is a useful technique when the fitness function cannot be exactly determined.

The flowchart of IGA is shown in Fig.7. We use a real coding to represent chromosomes. The chromosome has five bits and each bit corresponds to $w_1$, $w_2$, $w_3$, $w_4$ and $w_5$, respectively. A roulette wheel selection is used as a selection operator. A two points crossover is used to generate two children from two selected parents. In the two points crossover, two points are randomly selected and everything between the two points is swapped between the parent organisms. We use Eq.(7) for mutation.

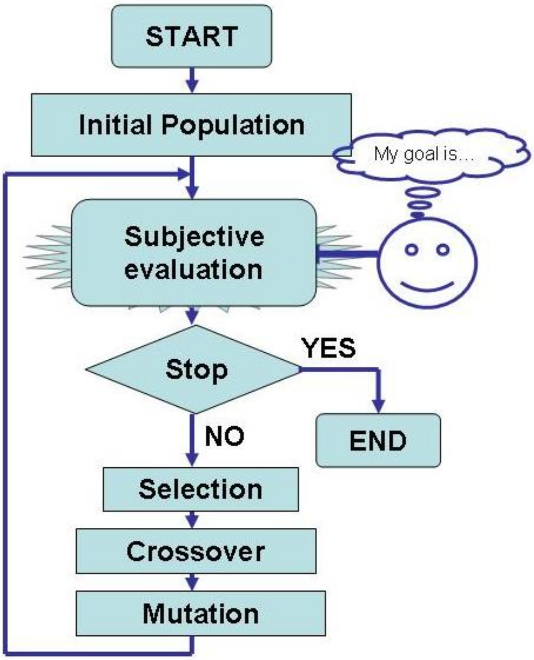


**Fig. 7.** Flowchart of the IGA

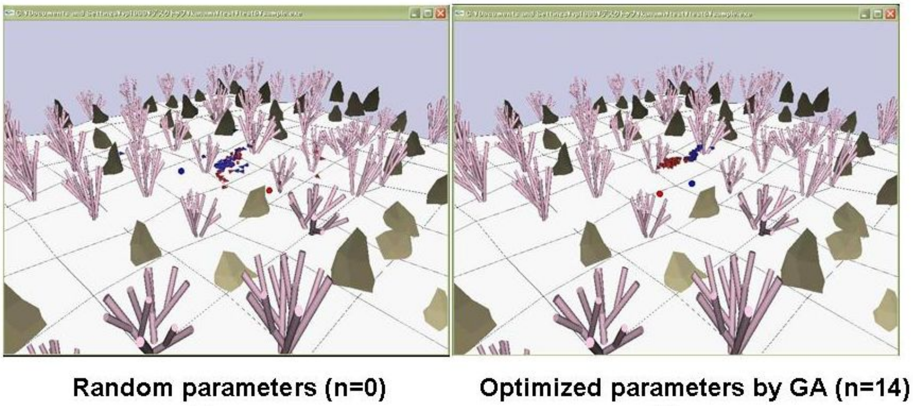
$$x' = x_l + \beta(x_u - x_j), \quad (7)$$

where $x_u$ and $x_l$ are upper limit and lower limit of the coefficients, respectively. β is a random value between 0 and 1. The chromosome and the bit for mutation are randomly selected.

Since it is difficult to give a mathematical fitness function to evaluate each individual (which is better and so on), we use an interactive method (which is also known as interactive GA) for evaluation. The fitness value of each individual is given by user's subjective evaluations. The user rates each individual (school) with 5-level score (1, 2, 3, 4, 5) based on his impression on behaviors of the fish school. Higher score, higher fitness. Compared with our previous classical GA[8], which used some mathematical fitness functions, the interactive GA can give us a better result.

# 4 Experimental Results

We have made an interactive artificial fish school system [9] based on the extended boids model and the system is made by Open GL. The examples of the system are shown in Fig.8. Two fish schools are simulated in the system. Figure 8(a) is a result with random parameters (which is also used as an initial value of GA; generation=0) and Fig.8(b) is a result with the classical GA-based optimization (after 14 generations). It can be seen that though the aggregate motions of fish schools was improved by using the GA-based optimization, but two schools have massed. The parameters are not really optimized because of the fitness function. The results by IGA are shown in Fig.9. Figure 9(a) is a result with random parameters (which is also used as an initial value of IGA; generation=0) and Fig.9(b) is a result with the proposed interactive GA-based optimization (after 14 generations). It can be seen that by using the interactive GA-based optimization, the aggregate motions of fish schools become more realistic and similar to behaviors of real fish world.

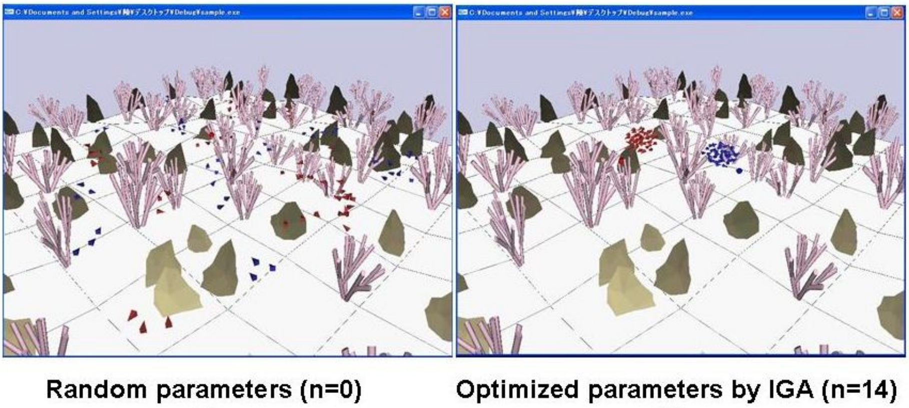


Random parameters (n=0)

Optimized parameters by GA (n=14)

(a) (b)

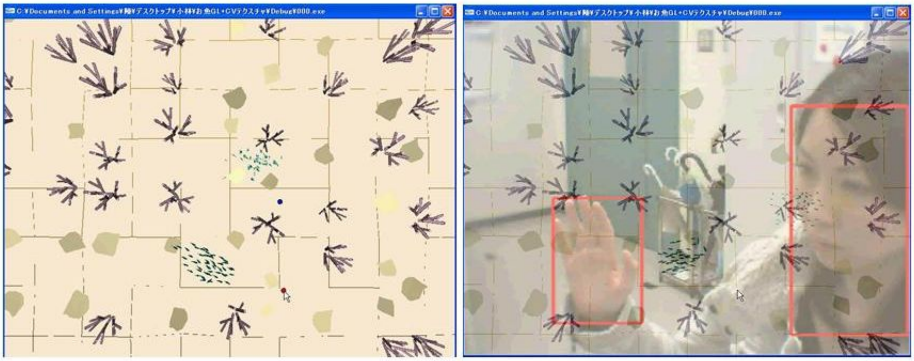
**Fig. 8.** (a) random parameters (generation=0); (b) with GA (generation=14)



(a) (b)

**Fig. 9.** (a) random parameters (generation=0); (b) with GA (generation=14)

We also added some interactive functions to the system. The user can interactive with the fish schools through a touch panel display or a web camera. Two examples are shown in Fig.10. Figure 10(a) is an example that the user can move the school by moving the feed. Figure 10(b) is another example that the web camera detects the hand or face of the user and the fish school will follow the movements of the user's hand or face.



(a) (b)

**Fig. 10.** Two examples of interaction between the user and the fish schools

## 5 Conclusions

In this paper, we proposed an extended boids model for simulating the aggregate moving of fish schools in a complex environment. Three behavior rules were added to the extended boids model: following a feed; avoiding obstacle; avoiding enemy boids.

We also proposed a genetic algorithm to optimize the coefficients. Experimental results showed that by using the GA-based optimization, the aggregate motions of fish schools become more realistic and similar to behaviors of real fish world.

**Acknowledgments.** This work was supported in part by the Strategic Information and Communications R&D Promotion Program (SCOPE) under the Grand No. 072311002 and the "Open Research Center" Project for Private Universities: matching fund subsidy from MEXT (Ministry of Education, Culture, Sports, Science, and Technology).

## References

1. Reynolds, C.W.: Flocks, Herds, and Schools: A Distributed Behavioral Model. Computer Graphics 21, 25–34 (1987)
2. DeAngelis, D.L., Shuter, B.J., Ridgeway, M.S., Blanchfield, P., Friesen, T., Morgan, G.E.: Modeling early life-history stages of smallmouth bass in Ontario lakes. Transaction of the American Fisheries Society, pp. 9–11 (1991)
3. Goldberg, D.: Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-Wesley, Reading (1989)
4. Forrest, S.: Genetic algorithms: principles of natural selection applied to computation. Science 261, 872–878 (1993)
5. Chen, Y.W., Nakao, Z., Arakaki, K., Fang, X., Tamura, S.: Restoration of Gray Images Based on a Genetic Algorithm with Laplacian Constraint. Fuzzy Sets and Systems 103, 285–293 (1999)
6. Chen, Y.W., Nakao, Z., Arakaki, K., Tamura, S.: Blind Deconvolution Based on Genetic Algorithms. IEICE Trans. Fundamentals E-80-A, 2603–2607 (1997)
7. Mendoza, N., Chen, Y.W., Nakao, Z., Adachi, T.: A hybrid optimization method using real-coded multi-parent EA, simplex and simulated annealing with applications in the resolution of overlapped signals. Applied Soft. Computing 1, 225–235 (2001)
8. Chen, Y.W., Kobayashi, K., Huang, Y., Nalao, Z.: Genetic Algorithms for Optimization of Boids Model. In: Gabrys, B., Howlett, R.J., Jain, L.C. (eds.) KES 2006. LNCS (LNAI), vol. 4252, pp. 55–62. Springer, Heidelberg (2006)
9. Kobayashi, K.: Interactive fish school generation system using GA. Graduation thesis of Ritsumeikan Univ. (2005)

# Classification of High-Resolution Satellite Images Using Supervised Locality Preserving Projections

Yen-Wei Chen[1,2] and Xian-Hua Han[1,2]

[1] Electronics & Information Eng., School ,Central South Univ. of Forestry and Tech.,
Changsha, 410004, China
[2] College of Information Science and Engineering, Ritsumeikan University, Japan
chen@is.ritsumei.ac.jp

**Abstract.** We proposed a new method based on supervised locality preserving projections (SLPP) for classification of high resolution satellite images. Compared with other subspace methods such as PCA and ICA, SLPP can preserve local geometric structure of data and enhance within-class local information. The proposed method has been successfully applied to IKONOS images and experimental results show that the proposed SLPP based method outperform ICA-based method. The proposed method can be practically incorporated into a GIS system.

**Keywords:** High-resolution satellite image, supervised, locality preserving projections, classification, multi-spectral image.

## 1 Introduction

Recently several high resolution satellites such as IKONOS, Quickbird have been launched and the high resolution images (1m) are available. A growing interest has been seen in geographic information system (GIS) constructions or updating based on high-resolution satellite images. The basic idea for GIS constructions is to extract the spatial information such as the road, the building and so on. The satellite image is a record of relative reflectance of particular wavelengths of electromagnetic radiation. Whether a particular target reflects specularly or diffusely, or somewhere in between, depends on the surface feature of the target and the wavelength of the incoming radiation. Multi-spectral information has been widely used for classification of remotely sensed images [1]. Since the spectra are combined by many factors such as object reflectance and instrumentation response, there are strong correlations among the spectra. Principal component analysis (PCA) has been proposed to reduce the redundancy among the spectra and find efficient representation for classifications or segmentations [2]. In our previous works, we proposed to apply independent component analysis (ICA) to learn the efficient spectral representation [3]. Since ICA features are higher-order uncorrelated while PCA features are second-order uncorrelated, higher classification performance has been achieved by ICA. Though ICA is a powerful method for finding efficient spectra representation, it is an unsupervised approach and it lacks the local geometric structure of data.

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 149–156, 2008.

Locality preserving projections (LPP) was proposed to approximate the eigenfunctions of the Laplace Beltrami operator on the image manifold, and be applied for face recognition and image indexing [4]. In this paper, we propose a new approach based on supervised locality preserving projections (SLPP) for classification of high-resolution satellite images. The scheme of the proposed method is shown in Fig.1. The observed multi-spectral images are first transformed by SLPP and then the transformed spectral components are used as features for classifications. A probabilistic neural network (PNN) [5] is used as a classifier. Compared with other subspace methods such as PCA and ICA, SLPP can not only find the manifold of images but also enhance the within-class local information. The proposed method has been successfully applied to IKONOS images and experimental results show that the proposed SLPP based method outperforms ICA-based method.

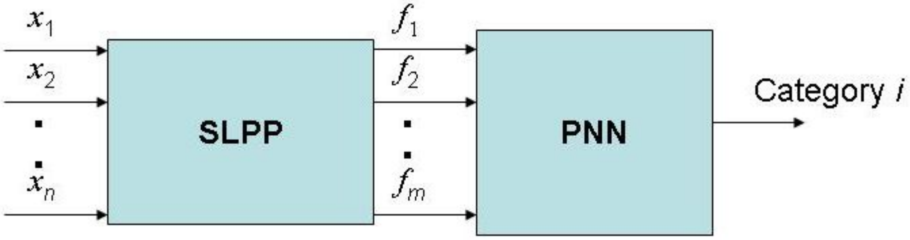


**Fig. 1.** The proposed method based on SLPP

The paper is organized as following: the supervised LPP for feature extractions is presented in Sec.2, the probabilistic neural network for classifications is presented in Sec.3 and the experimental results are shown in Sec.4. Finally, the conclusion is given in Sec.5.

This instruction file for Word users (there is a separate instruction file for LaTeX users) may be used as a template. Kindly send the final and checked Word and PDF files of your paper to the Contact Volume Editor. This is usually one of the organizers of the conference. You should make sure that the Word and the PDF files are identical and correct and that only one version of your paper is sent. It is not possible to update files at a later stage. Please note that we do not need the printed paper.

## 2 Supervised Locality Preserving Projections (SLPP)

The problem of subspace learning for image for feature extraction is the following. Given a set of spectral feature vectors $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_m$ in $\mathbf{R}^n$ of images, the goal is to find an efficient representation $\mathbf{f}_i$ of $\mathbf{x}_i$ such that $\left\|\mathbf{f}_i - \mathbf{f}_j\right\|$ reflects the neighborhood relationship between $\mathbf{f}_i$ and $\mathbf{f}_j$. In other word, if $\left\|\mathbf{f}_i - \mathbf{f}_j\right\|$ is small, then $\mathbf{x}_i$ and $\mathbf{x}_j$ are belong to same class. Here, we assume that the images reside on a sub-manifold embedded in the ambient space $\mathbf{R}^n$.

LPP seeks a linear transformation **P** to project high-dimensional data into a low-dimensional sub-manifold that preserves the local Structure of the data. Let $\mathbf{X}=[\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_m]$ denote the feature matrix whose column vectors is the sample feature vectors in $\mathbf{R}^n$. The linear transformation **P** can be obtained by solving the following minimization problem:

$$\min_{\mathbf{P}}\sum_{ij}(\mathbf{P}^T\mathbf{x}_i-\mathbf{P}^T\mathbf{x}_j)^2B_{ij} \tag{1}$$

where $B_{ij}$ evaluate the local structure of the image space. In this paper, we use normalized correlation coefficient of two samples as the penalty weight if the two samples belong to the same class:

$$B_{ij}=\begin{cases}\dfrac{\mathbf{x}_i^T\mathbf{x}_j}{\sum_{l=1}^n\mathbf{x}_{il}^2\sum_{l=1}^n\mathbf{x}_{jl}^2} & \text{if sample } i \text{ and } j \text{ are in same class}\\ 0 & \text{otherwise}\end{cases} \tag{2}$$

By simple algebra formulation, the objective function cam be reduced to:

$$\begin{aligned}&\frac{1}{2}\sum_{ij}(\mathbf{P}^T\mathbf{x}_i-\mathbf{P}^T\mathbf{x}_j)^2B_{ij}\\&=\sum_i\mathbf{P}^T\mathbf{x}_iD_{ii}\mathbf{P}^T\mathbf{x}_i-\sum_{ij}\mathbf{P}^T\mathbf{x}_iB_{ij}\mathbf{P}^T\mathbf{x}_j\\&=\mathbf{P}^T\mathbf{X}(\mathbf{D}-\mathbf{B})\mathbf{X}^T\mathbf{P}=\mathbf{P}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{P}\end{aligned} \tag{3}$$

**D** is a diagonal matrix; its entries are column (or row, since B is symmetric) sum of **B**, $D_{ii}=\sum_j B_{ij}$. **L=D-B** is the Laplacian matrix. Then, the linear transformation **P** can be obtained by minimizing the objective function under constraint:

$$\mathbf{P}=\underset{\mathbf{P}^T\mathbf{X}\mathbf{D}\mathbf{X}^T\mathbf{P}=1}{\arg\min}\ \mathbf{P}^T\mathbf{X}(\mathbf{D}-\mathbf{B})\mathbf{X}^T\mathbf{P} \tag{4}$$

Finally, the minimization problem can be converted to solving a generalized eigenvalue problem as follows:

$$\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{P}=\lambda\mathbf{X}\mathbf{D}\mathbf{X}^T\mathbf{P} \tag{5}$$

## 3 Probabilistic Neural Network (PNN)

The PNN model is based on Parzen's results on probability density function (PDF) estimators [5]. PNN is a three-layer feedforward network consisting of input layer, a pattern layer, and a summation or output layer as shown in Fig.2. We wish to form a Parzen estimate based on *K* patterns each of which is *n*-dimensional, randomly sampled from *c* classes. The PNN for this case consists of *n* input units comprising the input layer, where each unit is connected to one and only one of the category units. The connection from the input to pattern units represents modifiable weights, which

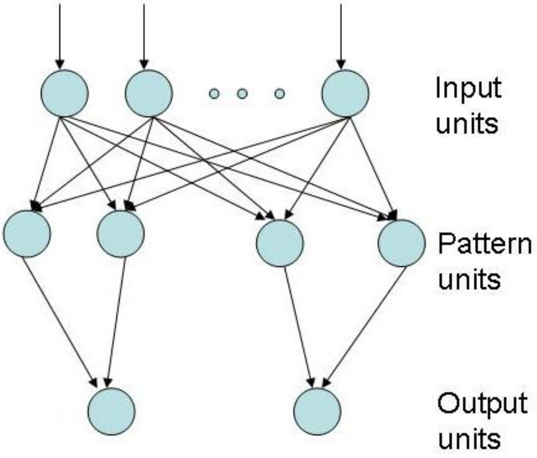


**Fig. 2.** PNN architecture

will be trained. Each category unit computes the sum of the pattern units connected to it. A radial basis function and a Gaussian activation are used for the pattern nodes.

The PNN is trained in the following way. First, each pattern (sample feature) **f** of the training set is normalized to have unit length. The first normalized training pattern is placed on the input units. The modifiable weights linking the input units and the first pattern unit are set such that $\mathbf{w}_1$=$\mathbf{f}_1$. Then, a single connection from the first pattern unit is connected to the category unit corresponding to the known class of that pattern. The process is repeated with each of the remaining training patterns, setting the weights to the successive pattern units such that $\mathbf{w}_k$=$\mathbf{f}_k$ for $k = 1,2,\cdots,K$. After such training we have a network which is fully connected between input and pattern units, and sparsely connected from pattern to category units. The trained

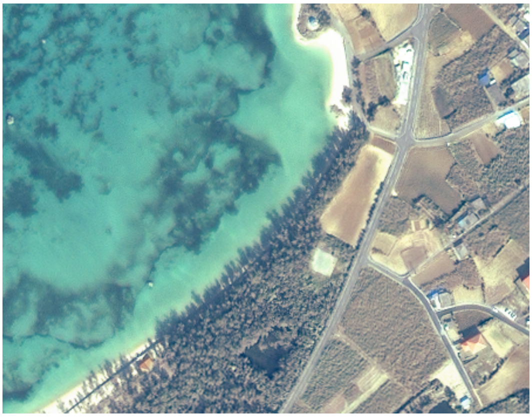


**Fig. 3.** IKONOS image (Copyright (C) 2003 Japan Space Imaging Corporation)

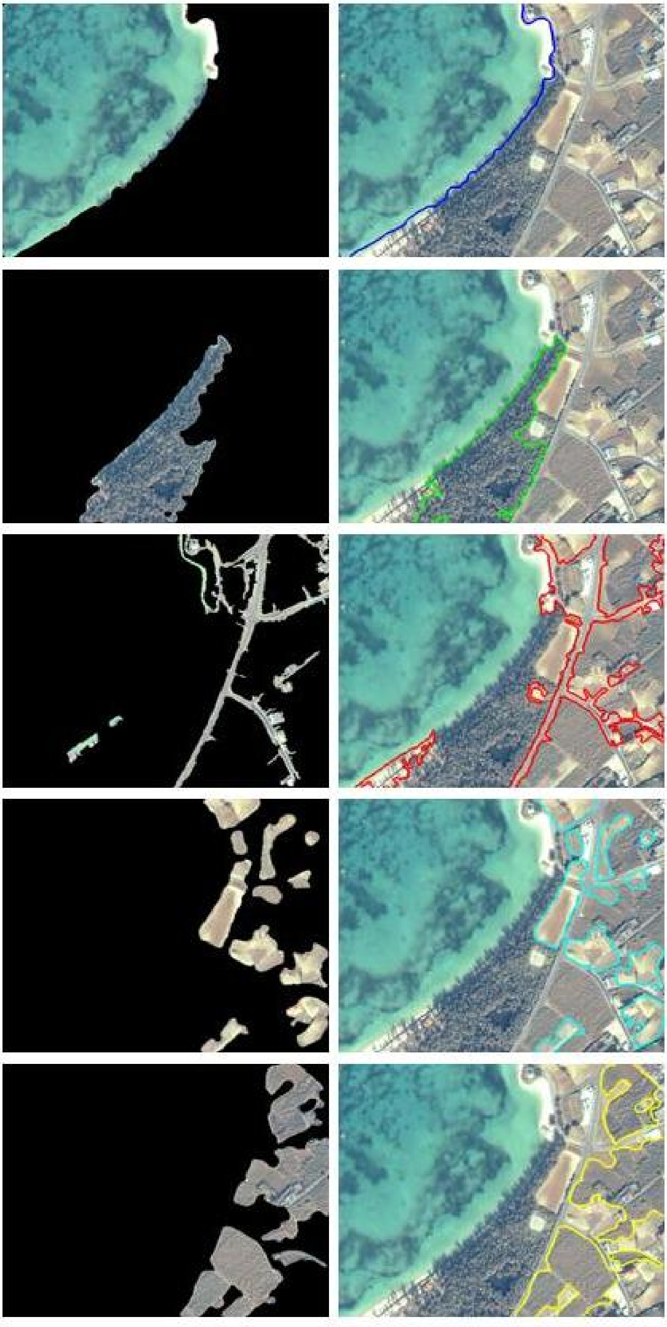


**Fig. 4.** Classification results by the proposed SLPP based method

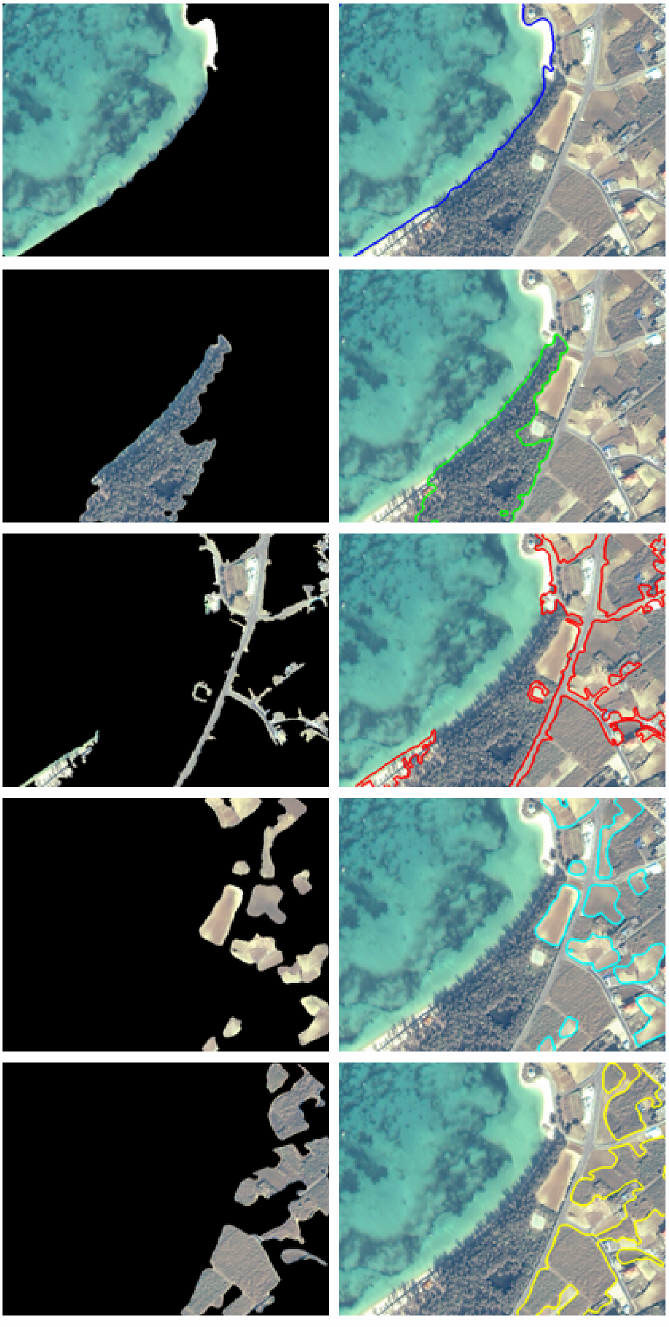

(a) (b)

**Fig. 5.** Classification results by ICA based method [3]

network is then used for classification in the following way. A normalized test pattern **f** is placed at the input units. Each pattern unit computes the inner product to yield the net activation **y**,

$$y_k = \mathbf{w}_k^T \cdot \mathbf{f} \tag{6}$$

and emits a nonlinear function of yk; each output unit sums the contributions from all pattern units connected to it. The activation function used is $\exp(\|\mathbf{x}-\mathbf{w}_k\|/\delta^2)$. Assuming that both **x** and **w**$_k$ are normalized to unit length, this is equivalent to using $\exp(\|\mathbf{x}-1\|/\delta^2)$.

## 4 Experimental Results

The proposed method has been applied to classification of IKONOS images. IKONOS simultaneously collects one-meter resolution black-and-white (panchromatic) images and four-meter resolution color (multi-spectral) images. The multi-spectral images consist of four bands in the blue (B), green (G), red (R) and near-infrared wavelength regions. And the multi-spectral images can be merged with panchromatic images of the same locations to produce "pan-sharpened color" images of 1-m resolution. In our experiments, we use only RGB spectral images for classifications. One typical IKONOS color image is shown in Fig.3.

The classification results by the proposed method are shown in Fig.4. The segmented sea, forest, road, ground, field are shown in Fig.4(a) and the extracted boundaries from each region are overlapped on the original image, which are shown in Fig.4(b). In order to make a comparison, the results by ICA are also shown in Fig.5. It can be seen that satisfactory classification results were obtained and the proposed SLPP based method outperforms ICA-based method. The proposed method can be practically incorporated into a GIS system. Detailed about GIS construction based on high-resolution satellite images has been published in [6].

## 5 Conclusions

In this paper, we proposed a new approach based on supervised locality preserving projections (SLPP) for classification of high-resolution satellite images. The observed multi-spectral images are first transformed by SLPP and then the transformed spectral components are used as features for classifications. A probabilistic neural network (PNN) is used as a classifier. Compared with other subspace methods such as PCA and ICA, SLPP can not only find the manifold of images but also enhance the within-class local information. The proposed method has been successfully applied to IKONOS images and experimental results show that the proposed SLPP based method outperforms ICA-based method.

**Acknowledgments.** This work was supported in part by the Strategic Information and Communications R&D Promotion Program (SCOPE) under the Grand No. 072311002.

## References

1. Avery, T.E., Berlin, G.L.: Fundamentals of Remote Sensing and Airphoto Interpretation. Macmillan Publishing Co., New York (1992)
2. Murai, H., Omatsu, S., OE, S.: Principal Component Analysis for Remotely Sensed Data Classified by Kohonen's Feature Mapping Preprocessor and Multi-Layered Neural Network Classifier. IEICE Trans.Commun. E78-B12, 1604–1610 (1995)
3. Zeng, X.-Y., Chen, Y.-W., Nakao, Z.: Classification of remotely sensed images using independent component analysis and spatial consistency. Journal of Advanced Computational Intelligence and Intelligent Informatics 8, 216–222 (2004)
4. He, X., Niyogi, P.: Locality Preserving Projections. In: Advances in Neural Information Processing Systems 16, Vancouver, Canada (2003)
5. Specht, Donald, F.: Enhancements to Probabilistic Neural Networks. In: Proceedings of the International Joint Conference on Neural Networks (IJCNN 1992), vol. 1, pp. 761–768 (1992)
6. Zaman, M.S., Chen, Y.-W., Zeng, X.-Y., Kang, D., Miyagi, H.: Construction of Graph-Structured GIS Road Map Using High-Resolution Satellite Image. Information 7(4), 527–536 (2004)

# Downsizing Multigenic Predictors of the Response to Preoperative Chemotherapy in Breast Cancer

René Natowicz[1], Roberto Incitti[2], Roman Rouzier[3], Arben Çela[1], Antõnio Braga[4],
Euler Horta[4], Thiago Rodrigues[5], and Marcelo Costa[4]

[1] Université Paris-Est, ESIEE-Paris, France
{r.natowicz,a.cela}@esiee.fr
[2] Institut Mondor de Médecine Moléculaire, Créteil, France
roberto.incitti@inserm.fr
[3] Hôpital Tenon, department of Gynecology, Paris, France
roman.rouzier@tnn.aphp.fr
[4] Universidade Federal de Minas Gerais, Depto. Engenharia Eletrônica, Brazil
{apbraga,eulerhorta}@cpdee.ufmg.br, macosta.est@gmail.com
[5] Universidade Federal de Lavras, Depto. Ciência da Computação, Brazil
thiago@dcc.ufla.br

**Abstract.** We present a method for designing efficient multigenic predictors with few probes and its application to the prediction of the response to preoperative chemotherapy in breast cancer.

In this study, each DNA probe was regarded as an elementary predictor of the response to the chemotherapy and the probes which were selected performed a faithful sampling of the training dataset.

In a first stage of the study, the prediction delivered by a multigenic predictor was that of the majority of the elementary predictions of its probes. For the data set at hand, the best majority decision predictor (MD predictor) had 30 probes. It significantly outperformed the best predictor designed on probes selected by p-value of a t-test (linear discriminant analysis on the 30 probes of least p-values).

In a second stage, the majority decision was replaced by a support vector machine (SVM) acting as a linear classifier. With the same set of probes, the performances of the SVM predictor were slightly better for both training and testing. Moreover, the performances of the best MD predictor were achieved with 43% less probes by SVM predictors (17 probes). This downsizing of the predictors is an interesting property for their potential use in clinical routine and for modeling the biological mechanisms underlying the patient's response to the chemotherapy.

## 1 Introduction

Nowadays, adjuvant and neoadjuvant (preoperative) administration of chemotherapy is based on prognostic factors, not on predictive ones. It is well known that the prognostic factors do not provide enough information for tailoring the treatment to the individual patients. Hence, nearly all breast cancer patients are given a standard chemotherapy treatment, despite their potentially poor response to the therapy, adverse side effects, and healthcare costs.

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 157–164, 2008.

The ability to predict the patients' response to the chemotherapy would be of high interest in the treatment of breast cancer for avoiding useless chemotherapy treatments and for selecting the most effective regimen for every patient. To this end, no single factor or biomarker ever has been in position to discriminate the patients who would respond to the treatment from those who would not. It appears that primary chemotherapy provides an ideal opportunity to correlate the gene expressions with the response to the treatment. Although gene expression microarrays provide novel tools and hold great promise in cancer research, the achievements in terms of improved prediction of drug sensitivity have been thus far rather moderate [1]. A strategy for translating microarray profiles into efficient clinical tests could consist in identifying small diagnostic gene-expression profiles with the help of microarrays then, in a second step, to validate the clinical usefulness of these genes, either retrospectively or prospectively, by making use of a simple and robust conventional assay, such as the quantitative reverse-transcriptase polymerase chain reaction (RT-PCR). Such a strategy requires microarrays analysis methods able to provide oncogenic signatures made out of few probe sets.

In the present study, every selected probe delivered an elementary prediction of the response to the treatment: pathologic complete response (*pcr*), residual disease (*nopcr*), or *unspecified*. In a first stage, we have defined very simple predictors whose predictions of the patient's response were that of the majority of its probes' elementary predictions: *PCR* if the majority of the probes individually predicted the response to be *pcr*; *NoPCR* if the majority was *nopcr*; and *UNSPECIFIED* in case of tie. In the second part of the study, the classification criterion of majority decision (MD) was replaced by support vector machines (SVM) acting as linear classifiers. The resulting classifiers had slightly better training and testing performances with the same numbers of probes. Moreover, the performances of the best MD predictor were achieved by a SVM predictor with significantly less probes, 17 probes instead of 30, i.e. more than 40% downsized.

In this paper, we will present the low level treatment through which a probe delivered an elementary prediction of the patient's response; the valuation function by which the probes were ranked then selected in this ranking; we will give the performances of the MD predictors then those of the SMV predictors for the dataset at hand.

## 2 Patients and Data

The clinical trial was conducted at the Nellie B. Connally Breast Center of The University of Texas M.D. Anderson Cancer Center [2]. One hundred thirty-three patients with stage I-III breast cancer were included. All patients underwent a single-pass, pretreatment fine-needle aspiration of the primary breast tumor before starting chemotherapy. Pretreatment gene expression profiling was performed with oligonucleotide microarrays (Affymetrix U133A) on fine-needle aspiration specimens. Patient cases were separated into patient *training cases* (82 cases) and patient *testing cases* (51 patient cases). At the completion of neoadjuvant chemotherapy, all patients had surgical resection of the tumor bed, with negative margins. Pathologic complete response (PCR) was defined as no histopathologic evidence of any residual invasive cancer cells in the breast, whereas residual disease was defined as any residual cancer cells after histopathologic study. The low level treatment of the microarray data was performed by software dCHIP V1.3

to generate probe level intensities. This program normalizes all arrays to one standard array that represents a chip with median overall intensity. Normalized gene expression values were transformed to the log10 scale for analysis.

The training set was composed of 82 patient data, each of which was the response to the treatment and the expression levels of the 22283 DNA probes. Among the training set, the response to the treatment was PCR for 21 patient cases and NoPCR for 61 cases.

The testing set was composed of 51 patient data among which the response to the treatment was PCR for 13 patient cases and NoPCR for 38 patients. Hence, the ratios of PCR to NoPCR patient cases were the same for both the training and testing datasets.

## 3 Probes Valuation

In order to put into light DNA probes conveying information on the two classes of patients, PCR and NoPCR, we chose to assign two sets of expression levels to each probe $s$, the sets $E_p(s)$ and $E_n(s)$, computed from the training data as follows [3,4]. Let $m_p(s)$ and $sd_p(s)$ be the mean and standard deviation of the expression levels of the probe $s$ for the PCR training cases, and let $m_n(s)$ and $sd_n(s)$ be those of the NoPCR training cases. The set of expression levels of the PCR training cases was defined as the set difference $E_p(s)$, $E_p(s) = [m_p(s) - sd_p(s), m_p(s) + sd_p(s)] \setminus [m_n(s) - sd_n(s), m_n(s) + sd_n(s)]$ and conversely for the NoPCR training cases, $E_n(s) = [m_n(s) - sd_n(s), m_n(s) + sd_n(s)] \setminus [m_p(s) - sd_p(s), m_p(s) + sd_p(s)]$.

*Discrete probes' predictions.* For any patient case, the individual prediction of a probe was a discrete value in the set $\{pcr,\ nopcr,\ unspecified\}$ : *pcr* if the expression level of patient $p$ was in the interval $E_p(s)$ and *nopcr* if it was in $E_n(s)$. Otherwise, the individual prediction value was *unspecified*.

*Probes' valuation function.* Let $p(s)$ be the number of PCR training cases correctly predicted *pcr* by the probe $s$, and let $n(s)$ be the number of the NoPCR training cases correctly predicted *nopcr* by the probe. The valuation function of the probes was defined so as to favor probes which correctly predicted high numbers of training cases and whose sets of correctly predicted training cases were 'good' samplings of the training set. To this end, we have considered the sensitivity and specificity values of the probe $s$, i.e. the ratios $p(s)/P$ and $n(s)/N$ of correctly predicted training cases. The valuation function $v(s)$, $v(s) \in [0, 1]$, was defined as $v(s) = 0.5 \times \left(\frac{p(s)}{P} + \frac{n(s)}{N}\right)$.

The figure 1 is the box-plot of the expression levels of a DNA probe of the gene BTG3, for the patients of the training set. This probe was one of the two equally top ranked probes, (cf. table 1.) From this figure, one can see that, given that the expression level of this probe was high, there was a high probability for the patient to have a PCR and symmetrically, given that the expression level was low, a high probability of residual disease. Hence, this probe delivered an information about both PCR and NoPCR classes. Up to a high rank, the probes selected by decreasing values conveyed information on both classes. In the ranking of the valuation function, the first *mono-informative* probe (giving information about only one class of patients) was at rank 63 (probe 207067_s_at of the gene HDC.) In contrast, the probe of smallest p-value was 203929_s_at, a probe of the gene MAPT. The box-plot of its expression levels

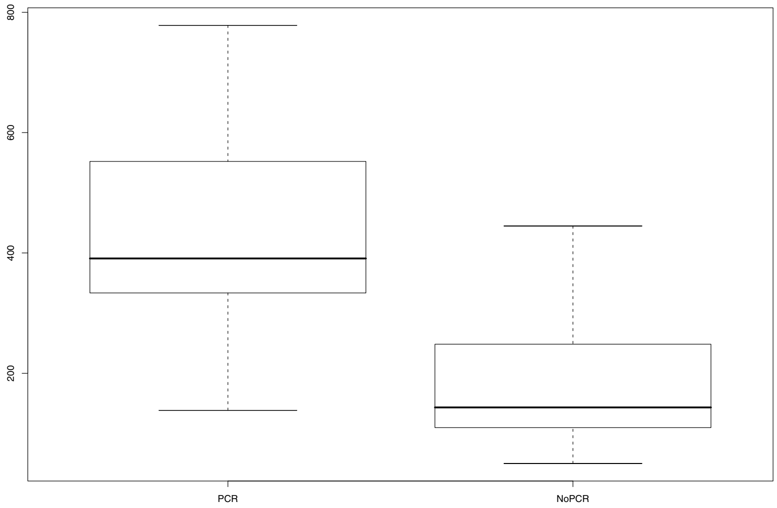


**Fig. 1.** Probe 205548_s_at of the gene BTG3, top ranked for the valuation function $v(s)$: box-plot of the expression levels

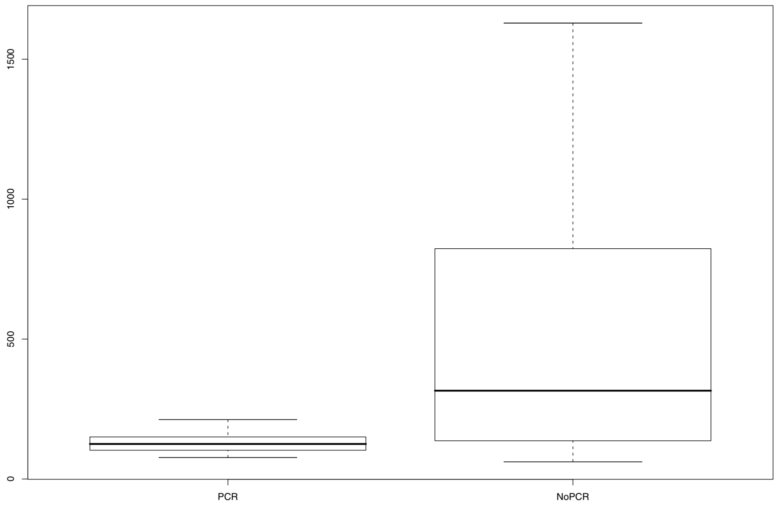


**Fig. 2.** Probe 203929_s_at of the gene MAPT, probe of smallest p-value to a t-test

(figure 2) shows that given a high expression level, the PCR probability was high. It also shows that a low expression level did not provide any information on the patient's class: according to the expression level of the probe, the probability to predict a PCR was that of a random choice with probability $\frac{P}{P+N}$, where $P$ and $N$ are the respective numbers of PCR and NoPCR cases of the training set.

**Table 1.** 33 top ranked genes. Gene: gene name in Hugo Gene nomenclature; probe: reference of the Affymetrix DNA probe set; $v(s)$: probe valuation; $p(s)$, $n(s)$: numbers of correct pcr and nopcr predictions for the 21 PCR and 61 NoPCR cases of the training set.

| Gene | Probe | $v(s)$ | $p(s)$ | $n(s)$ |
|---|---|---|---|---|
| BTG3 | 213134_x_at | 0.61 | 12 | 40 |
| BTG3 | 205548_s_at | 0.61 | 12 | 40 |
| GATA3 | 209604_s_at | 0.59 | 15 | 29 |
| GATA3 | 209603_at | 0.49 | 12 | 26 |
| THRAP2 | 212207_at | 0.46 | 8 | 34 |
| SCCPDH | 201826_s_at | 0.46 | 12 | 22 |
| SIL | 205339_at | 0.45 | 10 | 27 |
| KRT7 | 209016_s_at | 0.45 | 6 | 38 |
| MCM5 | 201755_at | 0.45 | 7 | 35 |
| NME3 | 204862_s_at | 0.44 | 10 | 25 |
| METRN | 219051_x_at | 0.44 | 11 | 22 |
| PDE4B | 211302_s_at | 0.43 | 9 | 27 |
| PHF15 | 212660_at | 0.42 | 7 | 32 |
| SSR1 | 200891_s_at | 0.42 | 7 | 32 |
| PISD | 202392_s_at | 0.42 | 11 | 20 |
| MELK | 204825_at | 0.41 | 8 | 28 |
| CA12 | 215867_x_at | 0.41 | 10 | 22 |
| CA12 | 214164_x_at | 0.41 | 10 | 22 |
| MAPK3 | 212046_x_at | 0.41 | 10 | 22 |
| GATA3 | 209602_s_at | 0.41 | 13 | 13 |
| BBS4 | 212745_s_at | 0.41 | 3 | 42 |
| DAPK1 | 203139_at | 0.41 | 9 | 24 |
| SAS | 203226_s_at | 0.40 | 7 | 29 |
| FLJ10916 | 219044_at | 0.40 | 8 | 26 |
| E2F3 | 203693_s_at | 0.40 | 8 | 26 |
| AHNAK | 220016_at | 0.40 | 9 | 23 |
| KLHDC3 | 214383_x_at | 0.40 | 9 | 23 |
| SFRS12 | 212721_at | 0.40 | 9 | 23 |
| SRPK1 | 202200_s_at | 0.39 | 6 | 31 |
| CXCR4 | 217028_at | 0.39 | 8 | 25 |
| KIF3A | 213623_at | 0.39 | 8 | 25 |
| MGC4771 | 210723_x_at | 0.39 | 8 | 25 |
| C11orf15 | 218065_s_at | 0.39 | 9 | 22 |

## 4 Multigenic Predictors with Majority Decision

We have defined the $k$-probes majority decision predictor (MD predictor) as the $k$ top ranked probes for the valuation function $v(s)$ together with the classification criterion of majority decision: for each patient case, the prediction was PCR if the number of elementary *pcr* predictions was strictly greater than the number of elementary *nopcr* ones, the prediction was NoPCR for the converse situation, and UNSPECIFIED in case of tie. In figure 3 are the training and testing accuracies of the first 41 MD $k$-predictors $(0 \leq k \leq 40.)$

The MD predictor of highest testing accuracy was the 33-probes predictor: accuracy=0.88 (2 FN and 4 FP), sensitivity=0.85, specificity=0.89, negative predictive value=0.944. The MD predictors of highest negative predictive value had $k = 27$, 29 and 30 probes: accuracy = 086 (1 FN and 6 FP), sensitivity = 0.92, specificity = 0.84, negative predictive value = 0.970.

The MD predictor of highest training accuracy was the 38-probes predictor: accuracy = 0.85 (3 FN, 9 FP), sensitivity = 0.86, specificity = 0.85, negative predictive value = 0.945. For $k = 27$ and 29 probes, the training performances were: accuracy = 0.83 (4 FN, 10 FP), sensitivity = 0.81, specificity = 0.84, negative predictive value = 0.927. For $k = 30$ probes the performances were: accuracy = 0.84 (4 FN, 9 FP), sensitivity = 0.81, specificity = 0.85, negative predictive value = 0.929. For $k = 38$ probes, the MD testing performances were: accuracy = 0.84 (5 FN, 3 FP), sensitivity = 0.76, specificity = 0.87, negative predictive value = 0.917.

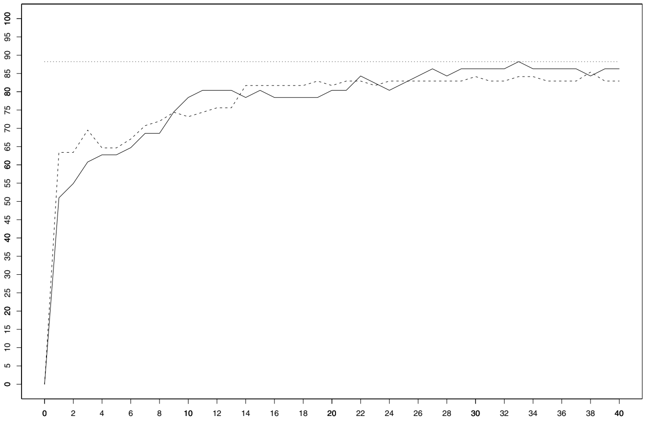


**Fig. 3.** Training and testing accuracies of the majority decision predictors. Solid line: testing accuracy, dashed line: training accuracy, dotted line: maximum testing accuracy (0.88) for $k = 33$ probes. $X$ axis: number of probes.

## 5 Training Set Sampling

Because the criterion the most widely used for selecting DNA probes in microarray studies for cancer research is the p-value of a t-test [1], and since the predictors designed with probes selected by our valuation function outperformed those designed with probes selected by the p-value [3], we were interested in finding a parameter which could be explicative of the observed difference of performances. It has appeared that the quality of the sampling performed by a given set of probes could account for the differences of performances.

The ratio of the numbers of PCR to NoPCR training cases of the data set at hand was $\frac{P}{N} = \frac{21}{61} = 0.34$. For the valuation $v(s)$, this ratio was in excellent agreement with that of the total numbers of *pcr* to *nopcr* correct predictions of the $k$ top ranked probes. For $20 \leq k \leq 50$ probes, the values of the ratios were between 0.33 and 0.34, and below 20 probes, the ratios were between 0.30 and 0.38. For the set of 30 top ranked probes, the mean number of correct predictions per probe was 35.16 and the ratio of the *pcr* to *nopcr* numbers of predictions was 0.34 (equal to the ratio of PCR to NoPCR numbers of cases in the training set).

For the p-value of a t-test, the set of 30 top ranked probes comprised 11 *mono-informative* probes. The mean number of correct predictions per probe was 37.67 (approximatively equal to that of the probes selected according to our probes valuation function), but the ratio of the *pcr* to *nopcr* numbers of predictions was far lesser, 0.14 this last ratio value being precisely equal to that of the whole set of probes.

From this, one could see that the NoPCR subset of training cases was over-sampled by the probes selected in the ranking of the p-value of a t-test, and more faithfully sampled by the probes selected according to the valuation function $v(s)$.

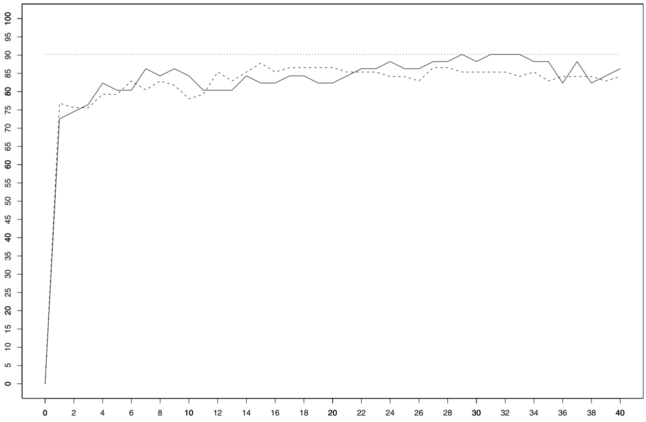


**Fig. 4.** Training and testing accuracies of the support vector machine used as linear classifiers. Solid line: testing accuracy, dashed line: training accuracy, dotted line: maximum training accuracy (0.90) for $k = 29$, 31,32,33 probes. $X$ axis: number of probes.

## 6 Multigenic Predictors with Support Vector Machine

The support vector machine $k$-predictor (SVM $k$-predictor) was defined as the $k$ top ranked probes, together with a linear classifier performed by an SVM trained on the learning set of patient cases. In figure 4 are the training and testing accuracies of the first 41 SVM $k$-predictors ($0 \leq k \leq 40$.) The maximum testing accuracy of the SVM $k$-predictors was achieved for $k = 29$, 31, 32 and 33 probes: accuracy=0.90 (1 FN and 4 FP). The sensitivity, specificity and negative predictive value of these four predictors were respectively 0.92, 0.89, 0.971. They improved the performances of the MD predictors of highest negative predictive value (cf. section 4 above.) From $k = 17$ to $k = 39$ the sensitivity was 0.92 (1 FN), the specificity ranged between 0.79 (8 FP) and 0.89 (4 FP), and the negative predictive value ranged from 0.969 to 0.971. The testing performances of the SVM 17-probes predictor were: accuracy=0.84 (1 FP, 7 FN), sensitivity=0.92, specificity=0.82, negative predictive value=0.969. These performances were very close to the testing performances of the MD 30-probes predictor. In particular, the negative predictive values were almost equal (31/32=0.967 for the SVM 17-probes vs. 32/33=0.970 for the MD 30-probes predictor.) Hence, the performances of the MD 30-probes predictor were achieved by the SVM predictor with 43% less probes. Furthermore, the SVM $k$-predictors showed slightly better training performances than the majority predictors (not reported here.)

## 7 Conclusion

With our approach of features selection, the DNA probes were regarded as elementary predictors of the response to the chemotherapy. We have presented a valuation function through which the probes behaving as faithful samplers of the training set were assigned high values. Two classifiers were evaluated for these probes: the non weighted majority decision among the elementary predictions of the probes, and a support vector machine performing a linear classification of the same elementary predictions. The SVM achieved the performances of the best MD predictor with 43% less probes. Its negative predictive and specificity values were respectively 0.92 and 0.89 (to be assessed by retrospective clinical studies.) Because of the high negative predictive value, such predictors could be of interest for supporting the decisions of not allocating patients to the treatment, and because of the high specificity value, 11% of the non responders would potentially be unadvisedly allocated to the treatment, instead of almost all of them with the nowadays systematic allocation.

The small number of probes involved in the SVM predictor could allow to design efficient predictors at very low cost, which is an important issue for their potential use in clinical routine and could be of great help for the task of modeling the biological mechanisms underlying the response to the chemotherapy.

**Acknowledgments.** The authors would like to thank the CAPES-COFECUB french-brazilian cooperation program and the CNPq for their support.

## References

1. Knudsen, S.: Cancer Diagnostics with DNA Microarrays. C.H.I.P.S edn. (2006)
2. Hess, K., Anderson, K., Symmans, W., et al.: Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer. Journal of Clinical Oncology 24(26), 4236–4244 (2006)
3. Natowicz, R., Incitti, R., Guimaraes Horta, E., et al.: Prediction of the outcome of preoperative chemotherapy in breast cancer by DNA probes that convey information on both complete and non complete responses. BMC Bioinformatics 9(149) (March 2008)
4. Natowicz, R., Braga, A., Incitti, R., et al.: A new method of dna probes selection and its use with multi-objective neural networks for predicting the outcome of breast cancer preoperative chemotherapy. In: European symposium on artificial neural networks, ESANN 2008 (May 2008)

# Classification of Sporadic and BRCA1 Ovarian Cancer Based on a Genome-Wide Study of Copy Number Variations

Anneleen Daemen[1,⋆], Olivier Gevaert[1], Karin Leunen[2], Vanessa Vanspauwen[3], Geneviève Michils[3], Eric Legius[3], Ignace Vergote[2], and Bart De Moor[1]

[1] Department of Electrical Engineering (ESAT)
Katholieke Universiteit Leuven, Leuven, Belgium
[2] Department of Obstetrics and Gynaecology, Division of Gynaecologic Oncology
Multidisciplinary Breast Centre, University Hospital Leuven, Leuven, Belgium
[3] Department of Human Genetics, University Hospital Leuven, Leuven, Belgium

**Abstract.** *Motivation:* Although studies have shown that genetic alterations are causally involved in numerous human diseases, still not much is known about the molecular mechanisms involved in sporadic and hereditary ovarian tumorigenesis.

*Methods:* Array comparative genomic hybridization (array CGH) was performed in 8 sporadic and 5 BRCA1 related ovarian cancer patients.

*Results:* Chromosomal regions characterizing each group of sporadic and BRCA1 related ovarian cancer were gathered using multiple sample hidden Markov Models (HMM). The differential regions were used as features for classification. Least Squares Support Vector Machines (LS-SVM), a supervised classification method, resulted in a leave-one-out accuracy of 84.6%, sensitivity of 100% and specificity of 75%.

*Conclusion:* The combination of multiple sample HMMs for the detection of copy number alterations with LS-SVM classifiers offers an improved methodological approach for classification based on copy number alterations. Additionally, this approach limits the chromosomal regions necessary to distinguish sporadic from hereditary ovarian cancer.

## 1 Introduction

Many defects in human development leading to e.g. cancer and mental retardation are due to gains and losses of chromosomes and chromosomal segments. These aberrations defined as regions of increased or decreased DNA copy number can be detected using an array comparative genomic hybridization (array CGH) technology. This technique measures variations in DNA copy number within the entire genome of a disease sample compared to a normal sample [1]. This makes array CGH ideally suitable for a genome-wide identification and

[⋆] Corresponding author.

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 165–172, 2008.

localization of genetic alterations involved in human diseases. An overview of algorithms for array CGH data analysis is given in [2]. Segmentation approaches identify adjacent clones with a same mean log ratio. These methods have as disadvantages that a further analysis is needed to determine the segments that are gained or lost and that results become unsatisfactory with high noise levels in the data. Therefore, segmentation and classification should be performed simultaneously because these two tasks can improve each other's performance. A popular method to combine them is the hidden Markov Model (HMM) with states defined as loss, neutral, one-gain and multiple-gain. Recently, this traditional procedure has been exploited to a multiple sample HMM in which a class of samples instead of individual samples is modeled by sharing information on copy number variations across multiple samples [3]. Here, we present a method to identify copy number alterations with the multiple sample HMM and that goes beyond the exploratory phase by using these alterations as features in a supervised classification setting.

For classification, we used the class of kernel methods which is powerful for pattern analysis. In recent years, these methods have become a standard tool in data analysis, computational statistics, and machine learning applications [4]. Their rapid uptake in bioinformatics is due to their reliability, accuracy and computational efficiency, which has been demonstrated in countless applications [5]. More specifically, as supervised classification algorithm we made use of the Least Squares Support Vector Machine (LS-SVM) which is an extension of the more regular SVM and has been developed in our research group by Suykens et al [6]. On high dimensional data, the LS-SVM is easier and faster compared to the SVM.

We applied our method on ovarian cancer which is the fourth most common cause of cancer death and ranks as the most frequent cause of death from gynaecological malignancies among women in western countries [7]. In a total of 5-10% of epithelial ovarian carcinomas, a family history of breast and ovarian cancer is noted with germline mutations in the tumour suppressor genes BRCA1 or BRCA2. A mutation of the BRCA1 gene cumulates the risk for ovarian carcinoma with 26-85% while a BRCA2 mutation increases the cumulative risk with 10% [8].

The outline of this article is as follows. In section 2, we describe the data set and the array CGH technology used for the analysis as well as the multiple sample HMM, the classifier and the feature selection method applied. In addition, the workflow of our proposed methodology is given in detail. In Section 3, we describe our results on ovarian cancer and finally, conclusions and future research directions are given in Section 4.

## 2 Materials and Methods

### 2.1 Patients and Data

Data from patients treated for ovarian cancer at the University Hospital of Leuven, Belgium were collected for participation at this study. All tumour samples were collected at the time of primary surgery. Only patients with similar clinical characteristics were retained: eight sporadic and five BRCA1 related ovarian cancer patients. One patient with BRCA2 was excluded and none of the patients out of the sporadic group had a positive family history of breast and/or ovarian cancer. Array comparative genomic hybridization was performed using a 1Mb array CGH platform, version CGH-SANGER 3K 7 developed by the Flanders Institute for Biotechnology (VIB), Department of Microarray Facility, Leuven, Belgium.

### 2.2 Array Comparative Genomic Hybridization

Array comparative genomic hybridization (array CGH) is a high-throughput technique for measuring variations in DNA copy number within the entire genome of a disease sample relative to a normal sample [1]. In an array CGH experiment, total genomic DNA from tumour and normal reference cell populations are isolated, different fluorescently labeled and hybridized to several thousands of probes on a glass slide. This allows to calculate the log ratios of the fluorescence intensities of the tumour to that of the normal reference DNA. Because the reference cell population is normal, an increase or decrease in the log intensity ratio indicates a DNA copy number variation in the genome of the tumour cells such that negative log ratios correspond to deletions (losses), positive log ratios to gains or amplifications and zero log ratios to neutral regions in which no change occurred.

### 2.3 Multiple Sample HMM

As was stated in the introduction, we will use a multiple sample hidden Markov Model (HMM) proposed by Shah et al [3] for the identification of chromosomal aberrations and to detect extended chromosomal regions of altered copy numbers labeled as gain or loss. The goal of this model is to construct features that distinguish the sporadic from the BRCA1 related group and subsequently to use them in a classifier (see Section 2.4). Because of the sensitivity of traditional HMMs to outliers being measurement noise, mislabeling and copy number polymorphisms in the normal human population, a robust HMM was first proposed by Shah et al [9] which handles outliers and integrates prior knowledge about copy number polymorphisms into the analysis. To further reduce the influence of various sources of noise on the detection of recurrent copy number alterations, Shah et al extended the robust HMM to a multiple sample version in which array CGH experiments from a cohort of individuals are used to borrow statistical strength across samples instead of modeling each sample individually [3]. This makes even copy number alterations in a small number of adjacent clones reliable when shared across many samples.

In this study, a multiple sample HMM is constructed on a chromosomal basis separately for the group of sporadic and the group of BRCA1 related ovarian cancer. Both HMMs result in chromosomal regions with genetic alterations characterizing sporadic and BRCA1 related samples, respectively. A differential region is defined as a chromosomal region which is gained/lost in one group while not being gained/lost in the other group.

### 2.4 Kernel Methods and Least Squares Support Vector Machines

The differential regions we just constructed are used as features in a classifier for which we chose kernel methods. These methods are a group of algorithms that do not depend on the nature of the data because they represent data entities through a set of pairwise comparisons called the kernel matrix [10]. This matrix can be geometrically expressed as a transformation of each data point $x$ to a high dimensional feature space with the mapping function $\Phi(x)$. By defining a kernel function $k(x_k, x_l)$ as the inner product $\langle \Phi(x_k), \Phi(x_l) \rangle$ of two data points $x_k$ and $x_l$, an explicit representation of $\Phi(x)$ in the feature space is not needed anymore. Any symmetric, positive semidefinite function is a valid kernel function, resulting in many possible kernels, e.g. linear, polynomial and diffusion kernels. In this manuscript, a linear kernel function was used.

An example of a kernel algorithm for supervised classification is the Support Vector Machine (SVM) developed by Vapnik [11] and others. Contrary to most other classification methods and due to the way data is represented through kernels, SVMs can tackle high dimensional data (e.g. microarray data). The SVM forms a linear discriminant boundary in feature space with maximum distance between samples of the two considered classes. This corresponds to a non-linear discriminant function in the original input space. This kernel method also contains regularization which allows tackling the problem of overfitting. We have shown that regularization seems to be very important when applying classification methods on high dimensional data [5]. A modified version of SVM, the Least Squares Support Vector Machine (LS-SVM), was developed by Suykens et al [6]. On high dimensional data sets, this modified version is much faster for classification because a linear system instead of a quadratic programming problem needs to be solved.

### 2.5 Feature Selection

Because it has been shown in [13] that univariate gene selection methods lead to good and stable performances across many cancer types and yield in many cases consistently better results than multivariate approaches, we used the method DEDS (Differential Expression via Distance Synthesis) [14]. This technique is based on the integration of different test statistics via a distance synthesis scheme because features highly ranked simultaneously by multiple measures are more likely to be differential expressed than features highly ranked by a single measure. The statistical tests which were combined are ordinary fold changes, ordinary t-statistics, SAM-statistics and moderated t-statistics. DEDS is available as a BioConductor package in R.

### 2.6 Proposed Methodology

Due to the limited number of samples, a leave-one-out (LOO) cross-validation strategy is applied. The 4 different steps that have to be accomplished in each LOO iteration are shown in Figure 1. After leaving out one sample, a multiple

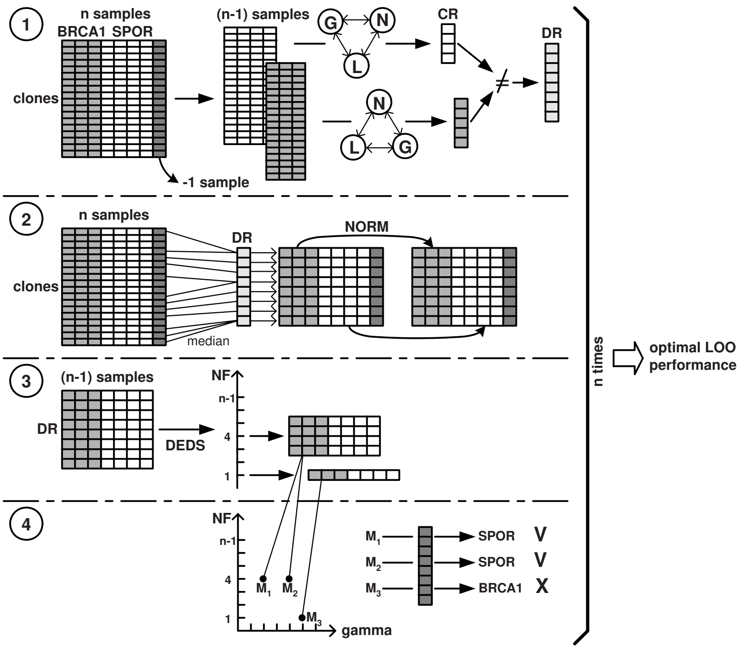


**Fig. 1.** Methodology consisting of 4 steps: step 1 - multiple sample HMM; step 2 - conversion of clones to differential regions and normalization per sample; step 3 - feature selection using DEDS; step 4 - LS-SVM training and validation on left out sample (CR = Chromosomal Region; DR = Differential Region; NORM = Normalization; DEDS = Differential Expression via Distance Synthesis; NF = Number of Features)

sample HMM (see Sect. 2.3) is constructed in step 1 for both groups of sporadic and BRCA1 related ovarian cancer to determine the chromosomal regions with genetic alterations that characterize each group. Combining these regions results in the chromosomal regions that are differential between the remaining n-1 sporadic and BRCA1 related samples. Because multiple clones can be located within each differential region, the clones need to be combined. This is done per sample in the second step by taking the median of the log ratios of the clones in each region. Afterwards, a standardization is performed per sample (i.e. meanshifting to 0 and autoscaling to 1) because the raw log ratios cannot be compared in absolute values between the samples. In step 3, DEDS determines which preprocessed log ratios, called features, best discriminate the n-1 samples (see Sect. 2.5). The number of included features is iteratively increased according to the obtained feature ranking without including more features than the number of samples on which the optimal number of features is determined [15]. This subset of features forms the input for classification in the last step (see Sect. 2.4). The LS-SVM contains a regularization parameter $\gamma$ which, together with the number of features needs to be optimized. For all possible combinations of $\gamma$ and number of features, an LS-SVM is built on the training set and validated on the left out sample. This is repeated n times such that each sample has been left out once. For the LS-SVM, a linear kernel function $k(x_k, x_l) = x_k^T x_l$ was chosen. An RBF kernel resulted in similar performances (data not shown).

## 3 Results

Our data set contains 8 sporadic and 5 BRCA1 related ovarian cancer patients. The array CGH data of chromosome 10 is shown in Figure 2 for 3 sporadic and 2 BRCA1 related samples. Both groups have a different profile within the first 3x$10^7$ base pairs and an amplification occurs within the BRCA1 related samples around 5x$10^7$ base pairs. When applying the proposed methodology on this data set, 11 out of 13 samples could be classified correctly using measured copy number changes in only 11 differential regions. The LS-SVM had a LOO accuracy of 84.6%, a sensitivity of 100% (5/5) and a specificity of 75% (6/8).

A comparison of the 11 differential regions found in each of the 13 LOO iterations shows a limited variability in the selected regions. Table 1 shows the number of LOO iterations in which the same features were chosen as the ones most differentially between all 13 samples. The top 5 of features with the lowest p-value according to DEDS appeared in 8 to 11 of the 13 LOO iterations. Three less significantly features appeared in 4 LOO iterations. These results strengthen our confidence that the chromosomal regions found with our methodology are robust and we hypothesize that genes in these regions participate in processes that distinguish sporadic from hereditary ovarian cancer.

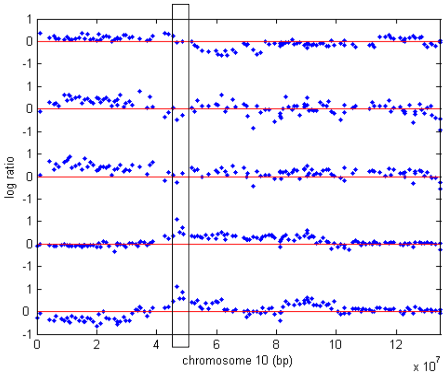


**Fig. 2.** Array CGH profile of chromosome 10 for 3 sporadic (top) and 2 BRCA1 related samples (bottom). The horizontal lines indicate the 0 log ratios for all samples. The vertical box indicates the amplification for the 2 BRCA1 related samples.

**Table 1.** Number of LOO iterations in which each of the 11 chromosomal regions was selected

| Feature | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Nb LOO iterations | 8 | 11 | 9 | 11 | 10 | 5 | 7 | 4 | 4 | 4 | 6 |

## 4 Conclusion and Future Work

In this manuscript, a new methodology is proposed in which copy number variations resulting from array CGH are transformed into features for classification purpose. This general method which is applicable to all types of cancer allows to find a small set of chromosomal regions for distinguishing two classes of patients and may further improve biological validation. It can also result in clinical relevant models for a simpler prediction based on a limited set of features. As increasing amounts of array CGH data become available, there is a need for algorithms to identify gains and losses statistically, rather than merely detect trends in the data. A large number of approaches for the analysis of array CGH data has already been proposed recently, ranging from mixture models and HMMs to wavelets and genetic algorithms [2]. However, most studies of cancer with gathered array CGH data apply less sophisticated methods for an exploratory analysis. Such studies apply a fixed threshold for defining gains and losses. A HMM on the contrary is a more intelligent way to detect copy number alterations in the genome of each sample by exploiting the spatial correlation between clones within an aberrated region. This makes the HMM also more robust against outliers such as measurement noise and wrongly recordings of locations of clones. Secondly, a robust HMM improves the reliability of the found chromosomal regions by taking into account copy number polymorphisms occurring in the normal human population. Thirdly, the multiple sample HMM improves the ability of detecting aberrations common for one group by borrowing strength across samples instead of modeling each sample individually. This makes also copy number alterations in a small number of adjacent clones reliable when shared across many samples and may prevent the loss of these possibly important biological features. Subsequently, the aberrations that are different between the group of sporadic and BRCA1 related samples are considered as features characterizing these samples. Finally, classification is performed to determine a small set of chromosomal regions that can distinguish sporadic from BRCA1 related ovarian cancer.

In the near future, an extensive study of the 11 differential regions may result in an increased knowledge on genes and pathways involved in sporadic versus hereditary ovarian cancer. Furthermore, we will analyze new patients with an in-house developed array CGH technology with a higher resolution to strengthen our hypotheses and to refine the found regions of genetic alterations possibly involved in ovarian cancer.

## Acknowledgements

AD is research assistant of the Fund for Scientific Research - Flanders (FWO-Vlaanderen). BDM is a full professor at the Katholieke Universiteit Leuven, Belgium. This work is partially supported by: **1.** Research Council KUL: GOA AMBioRICS, CoE EF/05/007 SymBioSys, PROMETA. **2.** Flemish Government: **a.** FWO projects G.0241.04, G.0499.04, G.0318.05, G.0302.07; **b.** IWT:

GBOU-McKnow-E, GBOU-ANA (biosensors), TAD-BioScope-IT, Silicos; SBO-BioFrame, SBO-MoKa, TBM-Endometriosis. **3.** Belgian Federal Science Policy Office: IUAP P6/25. **4.** EU-RTD: ERNSI, FP6-NoE Biopattern, FP6-IP e-Tumours, FP6-MC-EST Bioptrain, FP6-STREP Strokemap.

## References

1. Pinkel, D., Albertson, D.G.: Array comparative genomic hybridization and its applications in cancer. Nat. Genet. 37(Suppl.), 11–17 (2005)
2. Lai, W.R., Johnson, M.D., et al.: Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. Bioinformatics 21(19), 3763–3770 (2005)
3. Shah, S., Lam, W.L., et al.: Modeling recurrent DNA copy number alterations in array CGH data. Bioinformatics 23, i450–i458 (2007)
4. Shawe-Taylor, J., Cristianini, N.: Kernel methods for pattern analysis. Cambridge University Press, Cambridge (2004)
5. Pochet, N., De Smet, F., et al.: Systematic benchmarking of microarray data classification: assessing the role of nonlinearity and dimensionality reduction. Bioinformatics 20, 3185–3195 (2004)
6. Suykens, J.A.K., Van Gestel, T., et al.: Least Squares Support Vector Machines. World Scientific, Singapore (2002)
7. Gajewski, W., Legare, R.D.: Ovarian cancer. Surg. Oncol. Clin. N. Am. 7, 317–333 (1998)
8. Burke, W., Daly, M., et al.: Recommendations for follow-up care of individuals with an inherited predisposition to cancer. II. BRCA1 and BRCA2. Cancer Genetics Studies Consortium. J. Am. Med. Assoc. 277, 997–1003 (1997)
9. Shah, S., Xuan, X., et al.: Integrating copy number polymorphisms into array CGH analysis using a robust HMM. Bioinformatics 22(14), e431–e439 (2006)
10. Schölkopf, B., Tsuda, K., et al.: Kernel methods in computational biology. MIT Press, United States (2004)
11. Vapnik, V.: Statistical Learning Theory. Wiley, New York (1998)
12. Saeys, Y., Inza, I., et al.: A review of feature selection techniques in bioinformatics. Bioinformatics 23(19), 2507–2517 (2007)
13. Lai, C., Reinders, M.J.T., et al.: A comparison of univariate and multivariate gene selection techniques for classification of cancer datasets. BMC Bioinformatics 7, 235–244 (2006)
14. Yang, Y.H., Xiao, Y., et al.: Identifying differentially expressed genes from microarray experiments via statistic synthesis. Bioinformatics 21(7), 1084–1093 (2005)
15. Li, W., Yang, Y.: How many genes are needed for a discriminant microarray data analysis. In: Lin, S.M., Johnson, K.F. (eds.) Methods of Microarray Data Analysis, pp. 137–150. Kluwer Academic, Dordrecht (2002)

# Rule-Based Assistance to Brain Tumour Diagnosis Using LR-FIR

Àngela Nebot[1,⋆], Félix Castro[1], Alfredo Vellido[1], Margarida Julià-Sapé[2,3], and Carles Arús[3,2]

[1] Dept. de Llenguatges i Sistemes Informàtics - Universitat Politècnica de Catalunya
C. Jordi Girona, 1-3. 08034, Barcelona, Spain.
{angela,fcastro,avellido}@lsi.upc.edu
http://www.lsi.upc.edu/~websoco/AIDTumour
[2] Centro de Investigación Biomédica en Red en Bioingeniería, Biomateriales y Nanomedicina (CIBER-BBN), Cerdanyola del Vallès, Spain.
[3] Grup d'Aplicacions Biomèdiques de la RMN (GABRMN)
Departament de Bioquímica i Biología Molecular (BBM). Unitat de Biociències
Universitat Autònoma de Barcelona (UAB), Cerdanyola del Vallès, Spain
marga@carbon.uab.es,carles.arus@uab.es

**Abstract.** This paper describes a process of rule-extraction from a multi-centre brain tumour database consisting of nuclear magnetic resonance spectroscopic signals. The expert diagnosis of human brain tumours can benefit from computer-aided assistance, which has to be readily interpretable by clinicians. Interpretation can be achieved through rule extraction, which is here performed using the LR-FIR algorithm, a method based on fuzzy logic. The experimental results of the classification of three groups of tumours indicate in this study that just three spectral frequencies, out of the 195 from a range pre-selected by experts, are enough to represent, in a simple and intuitive manner, most of the knowledge required to discriminate these groups.

**Keywords:** Rule extraction, Fuzzy Inductive Reasoning, brain tumours, Magnetic Resonance Spectroscopy, Medical Decision Support Systems.

## 1 Introduction

Uncertainty is inherent to clinical oncology decision making, and poses a challenge for the development of any intelligent technologies with that purpose. The evidence, qualitative and quantitative, available to medical decision makers in oncology is growing exponentially. This situation justifies the design and development of computer-based decision support systems (DSS). The use of general

[⋆] The authors acknowledge funding from Spanish M.E.C. projects TIN2006-08114 and SAF2005-03650, as well as, from $1^{st}$ January 2003, Generalitat de Catalunya (grant CIRIT SGR2005-00863). Alfredo Vellido is a M.E.C. Ramón y Cajal researcher. Félix Castro is a research fellow within the PROMEP program of the Mexican Secretary of Public Education.

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 173–180, 2008.

medical DSS is now widespread and reasonably successful [1], but it is still rather uncommon to find any medical standard DSS using Computational Intelligence (CI) methods.

One of the potential drawbacks affecting the application of CI methods in general to the analysis of cancer data is the often limited interpretability of the results they yield. This is an extremely sensitive issue in a critical context such as oncology diagnosis. As stated in [2], "a DSS for medical diagnosis should support a comprehensible reasoning schema that corresponds to the human reasoning process". One way to overcome interpretability limitations, even though not the only one, is by explaining the operation of CI models using rule extraction methods. The interpretability of the model results should be greatly improved by their description in terms of reasonably simple and actionable rules that doctors and clinicians could rely on. In fact, rule extraction should provide clinicians, on whom the final responsibility for diagnosis rests, with an explanation about how a CI or related computer-based method has reached its decision [3].

Several authors have, in recent years, resorted to rule extraction from CI and related models in cancer research. Many of these involve the analysis of breast cancer data [4], although rule extraction for the classification of leukaemia and colon cancer data has also been proposed, for instance, in [5,6]. Given that this paper is mostly concerned with fuzzy methods, it is worth noting that fuzzy theory for cancer analysis has been applied in conjunction with evolutionary algorithms in [7], Artificial Neural Networks in [5,8] and rough sets in [9].

This paper describes a process of rule-extraction from a brain tumour database, and it is meant to be part of the design of the prototype medical DSS that is the goal of the AIDTumour (Artificial Intelligence Decision Tools for Tumour diagnosis [10]) research project. The multi-centre database under analysis in this study consists of Magnetic Resonance Spectroscopy (MRS) cases [11], corresponding to several tumour types. Rules to discriminate between tumours are extracted here using the Linguistic Rules in Fuzzy Inductive Reasoning (LR-FIR) algorithm [12]. At its core, FIR is a qualitative modeling and simulation methodology based on the observation of the input/output behavior of the system to be modeled, rather than on structural knowledge about its internal composition. LR-FIR is an extension that, starting from the systems' model that FIR identifies, is able to perform efficient generalization, deriving a set of actionable and realistic rules describing the systems' behavior.

The rest of the paper is organized as follows. The LR-FIR technique is described in section 2. The MRS dataset under study is briefly introduced in section 3. The rule extraction results are presented and discussed in section 4. The paper ends with a summary of conclusions and an outline of future research.

## 2 LR-FIR

The Fuzzy Inductive Reasoning (FIR) methodology is a mathematical tool for the modeling and simulation of complex systems. FIR is based on systems behavior rather that on structural knowledge. It is able to perform a selection

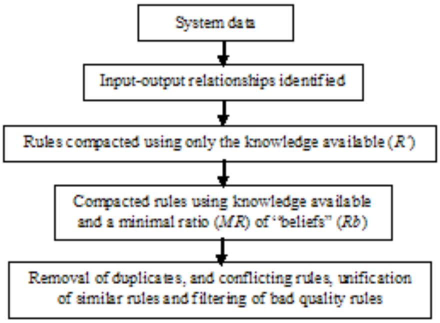

System data
Input-output relationships identified
Rules compacted using only the knowledge available (R')
Compacted rules using knowledge available and a minimal ratio (MR) of "beliefs" (Rb)
Removal of duplicates, and conflicting rules, unification of similar rules and filtering of bad quality rules

**Fig. 1.** Schematic representation of LR-FIR methodology

of the system's relevant variables and to obtain the causal and temporal relationships between them in order to infer the future behavior of that system. It also has the ability to describe systems that cannot easily be described by classical mathematics (e.g. differential equations), i.e. systems for which the underlying physical laws are not well understood. A FIR model is a qualitative, non-parametric, shallow model based on fuzzy logic, run under the Visual-FIR platform developed in Matlab®.

The FIR model consists of its structure (relevant variables) and a set of input/output relations (history behavior). Feature selection in FIR is based on the maximization of the models' forecasting power quantified by a Shannon entropy-based quality measure. Once the most relevant variables are identified, they are used to derive the set of input/output relations from the training data set. Detailed descriptions of the FIR methodology and Visual-FIR platform can be found in [13,14]. The history behavior is subsequently used as input for the LR-FIR technique, which extracts and compacts the information contained on it. The LR-FIR method aims to obtain interpretable, realistic and efficient rules, describing the behavior of the analyzed system. Fig. 1 shows its main phases.

The LR-FIR method can be summarized as a set of ordered steps:

1. *Basic compaction.* This is an iterative step that evaluates, one at a time, all the input/output relationships of a set $R$, which is compacted on the basis of the "knowledge" obtained by FIR. A subset $R_c$ can be compacted in the form of a single rule $r_c$, when all premises $P$ but one ($P_a$), as well as the consequence $C$, share the same values. Premises, in this context, represent the input features, whereas consequence is the output feature. If the subset contains all legal values $LV_a$ of $P_a$, all these relationships can be replaced by a single rule, $r_c$ , that has a value of -1 in the premise $P_a$. A -1 value means that this premise can take any of its possible values. When more than one -1 value, $P_{ni}$, is present in a compacted rule $r_c$, it is compulsory to evaluate the existence of conflicts by expanding all $P_{ni}$ to all their legal values $LV_a$, and comparing the resultant rules $X_r$ with the original relations $R$. If conflicts, $Cf$, exist, the compacted rule $r_c$ is rejected, and otherwise

accepted. In the latter case, the previous relationships subset, $R_c$ is replaced by the compacted rule $r_c$. Conflicts occur when one or more extended rules, $X_r$ have the same values in all its premises, $P$, but different values in the consequence $C$.

2. *Improved compaction.* Whereas the previous step only structures the available knowledge and represents it in a more compact form, the improved compaction step extends the knowledge base $R$ to cases that have not previously been used to build the model. Thus, whereas step 1 leads to a compacted data base that only contains knowledge, the enhanced algorithm contains undisputed knowledge and uncontested belief, $R_b$. The improved compaction is an extension of the basic compaction, where a consistent and reasonable minimal ratio, $MR$, of the legal values $LV_a$ should be present in the candidate subset $R_c$, to compact it in the form of a single rule $r_c$.

The obtained set of rules is subjected to a number of refinement steps: removal of duplicate rules and conflicting rules; unification of similar rules; evaluation of the obtained rules and removal of rules with low specificity and sensitivity values (this concepts are introduced later in the paper). For a more detailed description of LR-FIR methodology the user is referred to [12].

## 3 MRS Datasets

The analyzed data correspond to 217 single-voxel, short echo-time $^1$H-MR spectra acquired in vivo from brain tumour patients, classified according to the World Health Organization (WHO) system for diagnosing brain tumours by histopathological analysis of a biopsy sample. Three groups of tumours are analyzed in this study: G1, which includes 22 astrocytomas of grade 2, 6 oligoastrocytomas and 7 oligodendrogliomas, and is referred to as *low-grade gliomas*; G2, which includes 86 glioblastomas and 38 metastases, and is referred to as *high-grade malignant tumours*; and, finally, G3 includes 58 low-grade meningiomas. For each patient, there is a magnetic resonance spectrum consisting of 195 frequencies [11]. Each frequency (measured in parts per million (ppm), an adimensional unit of relative frequency position in the data vector) is treated as a data feature and, therefore, the dataset consists of 217 cases and 195 features. It was divided for analysis into training and test sets (balanced to account for class prevalence) containing, in turn, 163 and 54 patients.

## 4 Experimental Results and Discussion

FIR performs a selection of relevant variables in order to identify the structure of the model. In our experiments with the tumour groups described in the previous section, FIR found the frequencies 2.77ppm (herein referred to as f2.77), 2.33ppm (f2.33) and 1.34ppm (f1.34) to be the most relevant features. FIR also estimates the relative level of influence of the selected features on the discrimination of the tumour groups, which is 0.44 for f2.33, 0.43 for f1.34 and 0.13 for f2.77, (they add up to 1). In other words, f2.33 (corresponding to the presence of glutamate and macromolecules) and f1.34 (corresponding to lipids) are the most relevant frequencies in terms of the rules extracted for discriminating the tumour groups as described by the short echo-time MRS data. f2.77 (with no clear metabolic interpretation yet) also helps in the discrimination, but to a lesser extent and with a collateral role.

The three selected features are then used to compute the input/output relations, i.e. the history behavior that is the input information for the LR-FIR methodology. The rules obtained by LR-FIR for each of the three output classes, i.e. each group of brain tumours are presented in the first column of Table 1. A filtering threshold of 0.1 was applied, which means that the rules with specificity or sensitivity values lower or equal to 0.1 were deleted. Specificity is defined as one minus the ratio of the number of out-of-class data records that the rule identifies to the total number of out-of-class data. Sensitivity is the ratio of the number of in-class data that the rule identifies to the total number of in-class data. For simplicity, the antecedents of each rule are described in terms of numeric labels that correspond to intervals of values of the selected features, as described in Table 2. The second and third columns of Table 1 present the specificity and sensitivity metrics of the rules obtained from the training data set when applied to the test data set.

For illustration, the meaning of the first rule for G1 in Table 1 is as follows: if, for a given case, the value (L2-normalized height) of feature f2.33 is in the intervals with label 1 or 2, and the value of feature f1.34 is in the intervals with labels 1 or 2, then the case is inferred to be a *Low grade glioma* (G1).

Our goal from the onset was to obtain simple and interpretable models that represent as accurately as possible the behaviour of the system. The resulting rules described in Table 1 strike that balance: Only three frequencies of the spectra are enough to classify with an acceptable accuracy the three groups of

**Table 1.** Specificity and sensitivity metrics obtained for the test data set

| **Rules** | **Spec.** | **Sens.** |
|---|---|---|
| IF f2.33 IN 1-2 AND f1.34 IN 1-2 THEN CASE IN G1 | 0.73 | 0.92 |
| IF f2.77 IN 2-3 AND f2.33 IN 3 AND f1.34 IN 1 THEN CASE IN G1 | 0.85 | 0.077 |
| **JOINT METRICS *Low grade gliomas*** | **0.71** | **1** |
| IF f2.33 IN 1-2 AND f1.34 IN 2-3 THEN CASE IN G2 | 0.92 | 0.93 |
| **JOINT METRICS *High grade malignant*** | **0.92** | **0.93** |
| IF f2.33 IN 3 AND f1.34 IN 1-2 THEN CASE IN G3 | 0.98 | 0.92 |
| **JOINT METRICS *Meningiomas*** | **0.98** | **0.92** |

**Table 2.** Value intervals for each tumour group for variables: f2.77, f2.33 and f1.34

| | **f2.77** | **f2.33** | **f1.34** |
|---|---|---|---|
| Label 1 | [-2.143...2.2336] | [0.16822...4.6769] | [-2.1358...9.3643] |
| Label 2 | [2.2336...3.3045] | [4.6769...7.3388] | [9.3643...23.9711] |
| Label 3 | [3.3045...10.4054] | [7.3388...16.1629] | [23.9711...37.1991] |

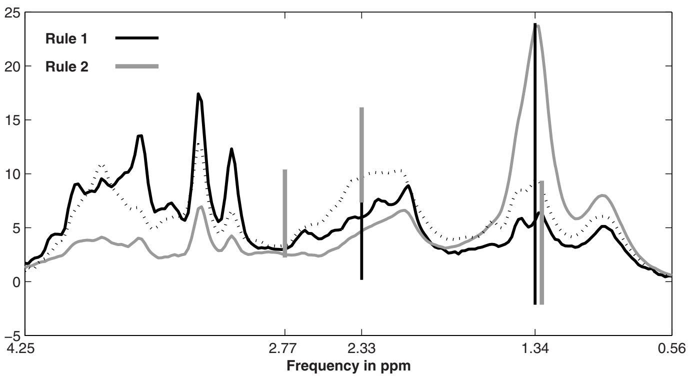


**Fig. 2.** Graphical representation, as vertical bars, of the rules described in Table 1 for *Low grade gliomas*, using the value intervals described in Table 2. They are described on top of the mean spectra for *Low grade gliomas* (G1, *solid black line*), *High grade malignant* (G2, *solid gray line*), and *Meningiomas*, (G3, *dotted black line*). *Vertical axis*: L2-normalized spectral intensity; *horizontal axis*: frequency chemical shift (ppm) with respect to water at 4.7 ppm.

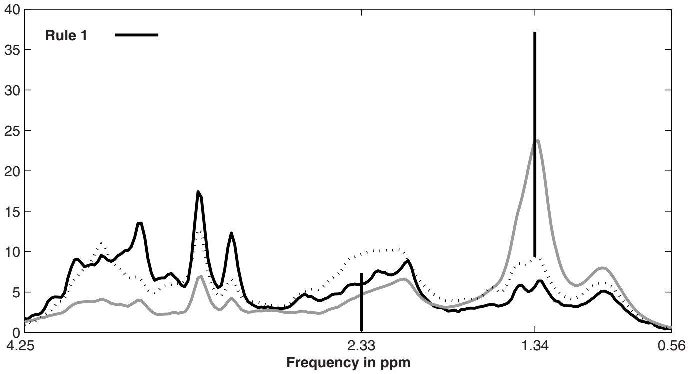


**Fig. 3.** Graphical representation, as in Fig. 2, of the rules described in Table 1 for *High grade malignant* tumours

tumours. Moreover, only two rules are needed to represent G1 quite accurately, while groups G2 and G3 only require a single rule for each one of them. This parsimonious representation of the three tumour groups under analysis should be extremely easy to act upon by medical experts. The interpretability of these rules can be further improved by representing them graphically on top of the characteristic spectra of the three types of tumours, as in Figs. 2, 3, and 4.

The visual interpretation of the rules for *high-grade malignant tumours* (G2) and *meningiomas* (G3) is straightforward. The case of *low-grade gliomas* (G1) is more striking, specially the value intervals of f2.77 and f2.33 for rule 2, which

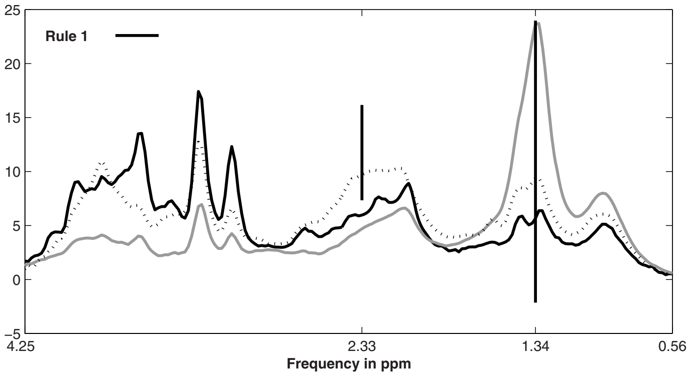


**Fig. 4.** Graphical representation, as in Fig. 2, of the rules described in Table 1 for *Meningiomas* tumour class

might seem to correspond to G3 instead of G1. In fact, this rule is only describing two outlier spectra of oligoastrocytomas (hence its low sensitivity). Therefore, this should be considered as a spurious rule. It illustrates two different things at once: the usefulness of rule visualization and the negative effect of data outliers on automated classification.

## 5 Conclusions

The interpretability of results is key in clinical oncology diagnostic assistance through computer-based methods. It has been shown in this paper, in a problem concerning the discrimination of diverse brain tumours using MRS data, that the interpretability of the problem can be greatly improved and simplified by its description in terms of a parsimonious set of simple and actionable rules that doctors and clinicians could rely on. The novel LR-FIR methodology has been used to this end in this study. Future research will focus on the integration of this rule extraction method and its results in the prototype medical DSS resulting from the AIDTumour [10] research project.

**Acknowledgments.** Authors gratefully acknowledge the former INTERPRET (EU-IST-1999-10310) European project partners. Data providers: Dr. C. Majós (IDI), Dr.À. Moreno-Torres (CDP), Dr. F.A. Howe and Prof. J. Griffiths (SGUL), Prof. A. Heerschap (RU), Dr. W. Gajewicz (MUL) and Dr. J. Calvar (FLENI); data curators: Dr. A.P. Candiota, Ms. T. Delgado, Ms. J. Martín, Mr. I. Olier and Mr. A. Pérez (all from GABRMN-UAB). The CIBER of Bioengineering, Biomaterials and Nanomedicine is an initiative of the "Instituto de Salud Carlos III" (ISCIII) of Spain.

## References

1. Garg, A.X., Adhikari, N.K.J., McDonald, H., Rosas-Arellano, M.P., Devereaux, P.J., Beyene, J., Sam, J., Haynes, R.B.: Effects of Computerized Clinical Decision Support Systems on Practitioner Performance and Patient Outcomes: A Systematic Review. J. Amer. Med. Assoc. 293, 1223–1238 (2005)
2. Tung, W.L., Quek, C.: GenSo-FDSS: a Neural-Fuzzy Decision Support System for Pediatric ALL Cancer Subtype Identification Using Gene Expression Data. Artif. Intell. Med. 33, 61–88 (2005)
3. Mitra, S.: Computational Intelligence in Bioinformatics. In: Peters, J.F., Skowron, A. (eds.) Transactions on Rough Sets III. LNCS, vol. 3400, pp. 134–152. Springer, Heidelberg (2005)
4. Vellido, A., Lisboa, P.J.G.: Neural Networks and Other Machine Learning Methods in Cancer Research. In: Sandoval, F., Gonzalez Prieto, A., Cabestany, J., Graña, M. (eds.) IWANN 2007. LNCS, vol. 4507, pp. 964–971. Springer, Heidelberg (2007)
5. Futschik, M.E., Reeve, A., Kasabov, N.: Evolving Connectionist Systems for Knowledge Discovery from Gene Expression Data of Cancer Tissue. Artif. Intell. Med. 28, 165–189 (2003)
6. Chen, Z., Lia, J., Wei, L.: A Multiple Kernel Support Vector Machine Scheme for Feature Selection and Rule Extraction from Gene Expression Data of Cancer Tissue. Artif. Intell. Med. 41, 161–175 (2007)
7. Peña-Reyes, C.A., Sipper, M.: A Fuzzy-Genetic Approach to Breast Cancer Diagnosis. Artif. Intell. Med. 17, 131–155 (1999)
8. Takahashi, H., Masuda, K., Ando, T., Kobayashi, T., Honda, H.: Prognostic Predictor with Multiple Fuzzy Neural Models Using Expression Profiles from DNA Microarray for Metastases of Breast Cancer. J. Biosci. Bioeng. 98, 193–199 (2004)
9. Hassanien, A.E.: Fuzzy Rough Sets Hybrid Scheme for Breast Cancer Detection. Image Vision Comput. 25, 172–183 (2007)
10. Artificial Intelligence Decision Tools for Tumour diagnosis (AIDTumour) research project, http://www.lsi.upc.edu/~websoco/AIDTumour
11. Julià-Sapé, M., et al.: A Multi-Centre, Web-Accessible and Quality Control-Checked Database of in Vivo MR Spectra of Brain Tumour Patients. Magn. Reson. Mater. Phy. MAGMA 19, 22–33 (2006)
12. Castro, F., Nebot, A.: Un Algoritmo para la Extracción Automática de Reglas Lógicas a partir de Modelos FIR. Technical Report, Universitat Politècnica de Catalunya LSI-07-7-R (2007)
13. Nebot, A., Cellier, F.E., Vallverdú, M.: Mixed Quantitative:Qualitative Modeling and Simulation of the Cardiovascular System. Comput. Meth. Prog. Bio. 55, 127–155 (1998)
14. Escobet, A., Nebot, A., Cellier, F.E.: Visual-FIR: A Tool for Model Identification and Prediction of Dynamical Complex Systems. Simul. Model. Pract. Th. 16, 76–92 (2008)
15. Vellido, A., Biganzoli, E., Lisboa, P.J.G.: Machine Learning in Cancer Research: Implications for Personalised Medicine. In: 16th European Symposium on Artificial Neural Networks (ESANN 2008). d-Side pub, Evere, Belgium (2008) (in press)

# Are Model-Based Clustering and Neural Clustering Consistent? A Case Study from Bioinformatics

Davide Bacciu[1,2], Elia Biganzoli[3], Paulo J.G. Lisboa[4], and Antonina Starita[2]

[1] IMT Lucca Institute for Advanced Studies,
P.zza San Ponziano 6, 55100 Lucca, Italy
d.bacciu@imtlucca.it
[2] Dipartimento di Informatica, Università di Pisa,
Largo B. Pontecorvo 3, 56127 Pisa, Italy
[3] Division of Medical Statistics and Biometry,
Istituto Nazionale per lo Studio e la Cura dei Tumori, Milan, Italy
[4] School of Computing and Mathematical Science,
Liverpool John Moores University, Liverpool, UK

**Abstract.** A novel neural network clustering algorithm, CoRe, is benchmarked against previously published results on a breast cancer data set and applying the method of Partition Around Medoids (PAM). The data serve to compare the samples partitions obtained with the neural network, PAM and model-based algorithms, namely Gaussian Mixture Model (GMM), Variational Bayesian Gaussian Mixture (VBG) and Variational Bayesian Mixtures with Splitting (VBS). It is found that CoRe, on the one hand, agrees with the previously published partitions; on the other hand, it supports the existence of a supplementary cluster that we hypothesize to be an additional tumor subgroup with respect to those previously identified by PAM.

## 1 Introduction

The unsupervised discovery of the processes underlying biological phenomena is an open challenge for the computational intelligence community that has been addressed by several unsupervised learning models with roots in statistics, fuzzy logic and neural networks, just to mention a few. In this work, we address the issue of unsupervisedly estimating the latent structure of a biomedical dataset, discovering data clusters as well as the most relevant sample covariates in each group. In particular, we compare a neural clustering algorithm, that is Competitive Repetitions Suppression learning (CoRe), with Bayesian approaches relating to Gaussian Mixtures Models. CoRe learning is a soft-competitive model inspired by a memory mechanism of the visual cortex, named Repetition Suppression (RS). The original CoRe model [1] has been extended recently in [2] to deal with high-dimensional data and with feature ranking. We apply this latter CoRe extension to analyze a case-study dataset from breast cancer research [3] with the intent of discovering clusters of functionally correlated samples, extracting

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 181–188, 2008.

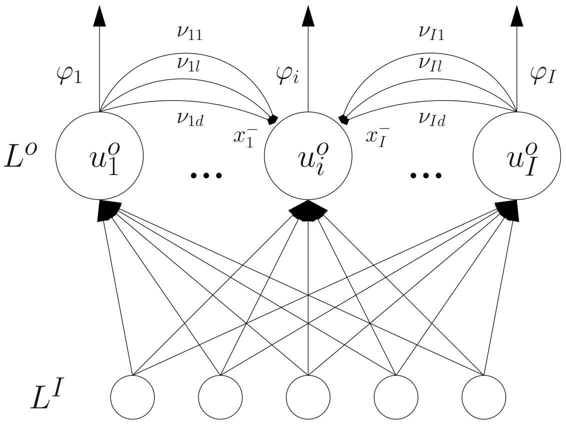


**Fig. 1.** CoRe neural network with lateral inhibition: intra-layer connections link the competitive neurons $u_i^o$ and propagate the Repetition Suppression. Lateral inhibition is applied selectively to the single input components $l$, i.e. there are $d$ lateral connections from neuron $u_j^o$ to neuron $u_i^0$, each weighted by a (possibly) different $\nu_{il}^t$.

cancer profiles characterized by different tumoral dynamics. Moreover, CoRe's feature ranking is used to gather insight into the markers that best describe the discovered bio-profiles.

Before reporting the experimental results, we introduce an alternative CoRe formulation that highlights its interpretation as a competitive neural network model, analyzing similarities and differences with unsupervised neural learning algorithms such as, for instance, SOM [4] and ART [5].

## 2 Neural Clustering with Competitive Repetition-Suppression Learning

A CoRe network is a two-layer neural network (see Fig. 1) where input nodes are fully connected to a layer of output units $L^o = \{u_1^o, \ldots, u_I^o\}$ that compete with each other through lateral connections. Each output unit $u_i^o$ is associated to a prototype vector $c_i \in \mathbb{R}^d$, determining its preferred stimulus, and to an activation function $\varphi_i(x_k)$ that determines the unit's response to the input pattern $x_k \in \mathbb{R}^d$. In other words, $c_i$ and $\varphi_i$ together determine the location and shape of the neuron's receptive field. In the remainder of the paper we will focus on univariate Gaussian activation functions $\varphi_i$ centered on $c_i$ and parametrical with respect to the spread . For each of the $l$-th components of the $d$ dimensional input vector, this defines an activation $\varphi_{il}(x_{kl}) \in [0, 1]$. This property will be used in the following to selectively suppress the irrelevant components in $c_i$.

The units in the outer layer are fully connected through a set a lateral inhibitory connections that serve to convey the suppressive potential to the loser neurons. More in detail, for an input pattern $x_k$ we first calculate the activation

$\varphi_i(x_k)$ of each output neuron $u_i^o$. Then, the most active units are selected to form the *winners pool*, i.e.

$$win_k = \{i \mid \varphi_i(x_k) \geq \theta_{win},\ u_i^o \in L^o\} \cup \{i \mid i = \arg\max_{j \in L^o} \varphi_j(x_k)\}, \tag{1}$$

while the remainder of the neurons is inserted into the *losers pool*, that is $lose_k = L^0 \setminus win_k$. Here $\theta_{win}$ determines the minimum activation level for winner units and, consequently, regulates the target selectivity of the neurons. The threshold $\theta_{win}$ does not directly determine the cluster number estimate, whereas it regulates the tradeoff between soft and hard competition among the neurons. Experimentally we have determined that the best results, for a generic clustering task, are obtained for $\theta_{win} \in [0.8, 1)$: lower $\theta_{win}$ values might produce slower convergence and incorrect cluster number estimation. Each unit $u_i^o$ generates an inhibitory output $x_{il}^-$ for each of the $l$ components of the $d$ dimensional input, that is

$$x_{il}^- = \begin{cases} \varphi_{il}(x_{kl})\nu_{il}^t \text{ if } i \in win_k \\ \qquad\qquad 0 \text{ if } i \in lose_k \end{cases} \tag{2}$$

where $0 \leq \nu_{il}^t \leq 1$ (i.e. the *stimulus predominance* in [2]) represents the weight of the inhibitory connection for the $l$-th component (see lateral links in Fig. 1). The inhibitory weight vector $\nu_i^t \in \mathbb{R}^d$ essentially measures the frequency of the patterns that are preferred by the $i$-th unit. The value of its $l$-th component (at time $t$) is computed as

$$\nu_{il}^t = \frac{1}{|\chi_t|} \sum_{x_k \in \chi_t} \min\left(\frac{\varphi_{il}(x_{kl})}{\varphi_{z(L^0,x_k)l}(x_{kl})}, 1\right) \tag{3}$$

where $\chi_t$ is the set of patterns $x_k$ presented to the network up to time $t$ and $z(L^0, x_k)$ returns the index of most active unit for the pattern $x_k$.

The inhibitory output is propagated to all the competing neurons through the lateral connections and is accumulated at each unit $i$ as $RS_{il}^t(x_k) = 1/|win_k| \sum_{j \in I_i^-} x_{jl}^-$, that is the $l$-th component of the *repetition suppression* generated at time $t$ for the pattern $x_k$. The term $|win_k|$ represents the cardinality of the winners pool for $x_k$ and is used to ensure $0 \leq RS_{il}^t \leq 1$, while $I_i^-$ is the set of the inhibiting connections for the $i$-th neuron. Finally, the $l$-th component of the prototype vector $c_i$ is updated as follows

$$\triangle c_{il}^t = \alpha_c \left[\delta_{ik} - (1 - \delta_{ik})\varphi_{il}(x_{kl})(RS_{il}^t(x_k))^2\right] \varphi_{il}(x_{kl})(x_{kl} - c_{il}^{t-1}) \tag{4}$$

where $\alpha_c$ is the learning rate, while $\delta_{ik}$ is the indicator function for the winners pool, that is $\delta_{ik} = 1$ if $i \in win_k$ and $\delta_{ik} = 0$ otherwise. Equation (4) essentially states that CoRe inhibition produces a penalization that is proportional to the frequency of the stimuli similar to the current input $x_k$. This penalization is applied only to the loser neurons, deflecting the prototype components $c_{il}^t$ from the respective $x_{kl}$ proportionally to the generated RS, while the winners' prototype is moved towards the current input vector.

The formulation in (4) can be used to compare CoRe to other competitive learning models in literature. Consider a typical prototype update function for a generic competitive neural network, that is $\triangle c_{il}^{t} = \alpha_c h(x_k, L)(x_{kl} - c_{il}^{t-1})$, where $h(x_k, L)$ is a function regulating the type of competition between the units in the layer $L$. In Self Organizing Maps [4] and Growing Neural Gas, $h()$ is a neighborhood function based on the proximity of the units to the best matching unit (BMU) in a given lattice, hence generating spatially ordered maps of the input stimuli, while in ART lateral inhibition serves to selectively shut-off committed neurons, allowing other nodes in the network to win the competition. In contrast, CoRe's $h()$ applies active negative reinforcement outside of the winners' pool. Moreover, the intra-layer links propagate a teaching signal that produces a long-term silencing of the neurons and a suppression of the irrelevant input components.

The prototype update rule in (4) can be obtained by differentiating the error

$$E_{il,k}^{t} = \delta_{ik}(1 - \varphi_{il}(x_{kl})) + (1 - \delta_{ik})\frac{1}{2}(\varphi_{il}(x_{kl})RS_{il}^{t}(x_k))^2 \qquad (5)$$

with respect to $c_{il}^{t}$ (see [2] for further details). The same approach can be taken for all the variables in the activation function. For instance, differentiating the Gaussian $\varphi_i$ with respect to the spread $\sigma_{il}$ leads to the following update rule

$$\triangle \sigma_{il}^{t} = \alpha_\sigma \left[\delta_{ik} - (1 - \delta_{ik})\varphi_{il}(x_{kl})(RS_{il}^{t}(x_k))^2\right] \varphi_i(x_{kl})\frac{(x_{kl} - c_{il})^2}{\sigma_{il}^3}. \qquad (6)$$

The spread adjustment process in (6) adaptively reduces the variance of winner units, thus making them more selective. This behavior, on the one hand, complies with the biological repetition suppression phenomenon and, on the other hand, ensures the convergence of the learning process.

The *relevance factor* for the $l$-th feature of the $i$-th neuron is modeled as

$$\hat{\nu}_{il}^{t} = \frac{1}{\nu_{il}^{t}|\chi_t|} \sum_{x_k \in win_{u_i^o}^{t}} \left\{\frac{\varphi_i(x_{kl})}{\varphi_{z(win_k, x_k)l}(x_{kl})}\right\}_{\mathbf{0}} \text{ s.t. } \{v\}_{\mathbf{0}} = \begin{cases} 0 \; v > 1 \\ v \; v \leq 1 \end{cases} \qquad (7)$$

where $win_{u_i^0}^{t}$ is the set of patterns $x_k \in \chi_t$ for which unit $u_i^0$ was in the winners pool, while the function $\{\cdot\}_{\mathbf{0}}$ flattens its argument to zero if it exceeds 1. The rationale behind this choice is to penalize the relevance of the $l$-th component of a prototype $c_i$ whenever it produces a high feature activation $\varphi_{il}(x_{kl})$ in correspondence with a low feature activation $\varphi_{jl}(x_{kl})$ in the unit $u_j^o$ that is the maximally active neuron for the pattern $x_k$, i.e. $j = z(win_k, x_k)$.

Based on the measure $\hat{\nu}_i^t \in \mathbb{R}^d$ CoRe defines a neuron silencing mechanism that resembles the unit commitment system of the ART model [5]. In brief, in ART networks uncommitted neurons can be activated whenever none of the committed units represents the current input stimulus sufficiently well (with respect to a given vigilance parameter). This process automatically draws from the pool of uncommitted, or dormant, neurons, to represent novel stimuli. In CoRe, all available neurons compete from the start for every stimulus. The relevance measure in (7) is then used to decide whether a neuron can be silenced or pruned from the network. The univariate components $\hat{\nu}_{il}^{t}$ of the relevance factor also identify which prototype dimensions $c_{il}$ best characterize the input patterns assigned to the $i$-th neuron.

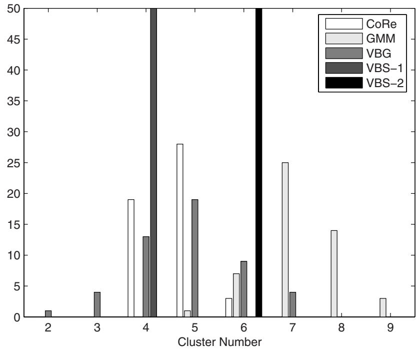


**Fig. 2.** Distribution of the cluster number estimate for the 50 independent runs

## 3 A Case Study in Breast Cancer Research

The learning algorithm described in the previous section can be readily applied to clustering tasks. In particular, each CoRe neuron $u_i^o$ can be interpreted as a cluster detector with centroid $c_i$: starting with an initially large neural population, CoRe iteratively prunes irrelevant neurons until it converges to an estimate of the number of clusters in the data. In this section, we study the performance of our neural approach on a recently published dataset [3] from breast cancer research. The dataset contains 633 samples consisting of five breast cancer biomarkers, that are ER (estrogen receptors), PR (progesterone receptors), Pro (Ki-67/MIB1 proliferation marker), NEU (HER2/NEU) and P53. CoRe is used to analyze the dataset and discover samples' sub-groups, possibly sharing a common bio-medical trait. Its results are compared with those obtained by three model-based algorithms, namely a Gaussian Mixture Model (GMM) [6] fitted using Expectation Maximization (EM) and two fully Bayesian model fitted by a variational approximation of the EM, that are Variational Bayesian Gaussian Mixture (VBG) [7] and Variational Bayesian Mixtures with Splitting (VBS) [8]. Both GMM and VBG estimate the number mixtures (i.e. clusters) by pruning those components with (almost) zero mixing weights. Conversely, VBS determines the number of mixtures by recursively splitting and pruning the existing components. Additionally, the results obtained by each algorithm are compared with the sample labeling discovered in previous work [3] by Partition Around Medoids (PAM) and k-means (KM). The results are based on 50 runs of each algorithm, with random initial prototype positions. Both GMM and VBM are initialized with 20 mixtures, VBS is initialized with one component and CoRe starts with 30 units (for CoRe meta-parameter settings refer to [2]).

First, we focus on determining the most likely number $K$ of sample groups in the data. Figure 2 shows the histogram of the cluster numbers estimated by the four algorithms during the 50 runs. The distribution of $K$ for GMM and VBG closely resembles a Gaussian centered on 7 and 5, respectively. CoRe and VBS produce sharper hypothesis: the former, in particular, suggests that the data can be grouped in 4 or 5 clusters, with few runs terminating with 6 clusters.

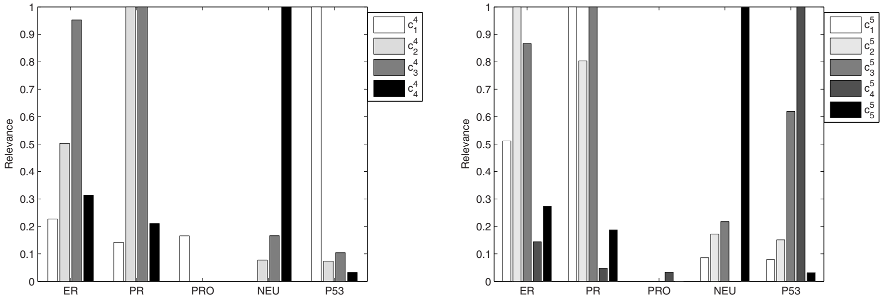
**Table 1.** Clustering concordance evaluated by $\kappa$ statistics: X's indicates when model comparison was not possible

| | CoRe | | | | VBG | | | VBS | | PAM |
|---|---|---|---|---|---|---|---|---|---|---|
| $K$ | KM | PAM | VBS | VBG | KM | PAM | VBS | KM | PAM | KM |
| $K = 4$ | 0.83 | 0.83 | 0.65 | 0.46 | 0.54 | 0.47 | 0.36 | 0.76 | 0.74 | 0.96 |
| $K = 5$ | 0.83 | X | 0.54 | 0.38 | 0.36 | X | 0.53 | 0.65 | X | X |
| $K = 6$ | 0.68 | X | 0.61 | X | 0.23 | X | 0.52 | 0.49 | X | X |

The behavior of VBS is peculiar: this algorithm behaves similarly to a wrapper approach that tries several configurations of the model until it finds a stable solution with respect to its internal criterion. Within this stability criterion, VBS always converged to a solution where $K = 6$ (see VBS-2 in Fig. 2). However, by taking a closer look at VBS learning dynamics one discovers that the maximum likelihood solution is obtained always for $K = 4$ (see VBS-1 in Fig. 2), while $K = 5$ is the second best-scoring solution with respect to the likelihood.

These results suggest that the most likely hypotheses are $K = \{4, 5, 6\}$. To evaluate the agreement between the data partitions produced by the different algorithms, we compared them with a baseline k-means clustering and with the sample labels discovered by PAM in [3]. Table 1 summarizes the concordance of the generated solutions in terms of the $\kappa$ statistics: the GMM model is not shown since it produced results only for $K = 6$, with a minimal overlap between its solutions and those produced by the other algorithms. Besides GMM, model based algorithms seem to produce partitions that are quite uncorrelated with respect to those generated by the other algorithms: only VBS shows a fair agreement with CoRe on the hypotheses $K = 4$ and $K = 6$. The $\kappa$ values in Table 1 show a substantial agreement between CoRe and k-means, especially for the hypotheses K=4 and K=5, that are those most strongly advised by CoRe. The concordance with the PAM labels is analyzed only on the $K = 4$ hypothesis since this is the cluster number estimated in [3]. The KM solution shows the highest agreement with PAM labels, confirming the results in [3], while Gaussian-based algorithms seem to find different solutions with respect to both KM and PAM. As a general comment, the model-based algorithms seem unable to converge to a shared sample classification, producing quite discording dataset partitions. CoRe, on the other hand, produces solutions that are a trade-off between the concordant sample classification produced by PAM and KM, and the alternative partition discovered by VBS.

To gather a better insight into the CoRe's cluster profiles we looked at the relevance $\hat{\nu}_{il}$ of the biomarkers characterizing CoRe's sample groups. Figure 3.a and 3.b show CoRe's feature relevance for the K=4 and K=5 models, respectively. Prolind (PRO) does not seem to play a significant role in the clustering process. The profiles in Fig. 3 also clearly show, in each partition, a single cluster differentiated by HER2/NEU expression, as expected.



(a) (b)

**Fig. 3.** CoRe relevance factors for the (a) four and (b) five clusters scenario

**Table 2.** Samples cross-distribution between PAM and CoRe ($K = 4$ and $K = 5$)

| | CoRe (K=4) | | | | CoRe (K=5) | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $PAM$ | $C_1^4$ | $C_2^4$ | $C_3^4$ | $C_4^4$ | $C_1^5$ | $C_2^5$ | $C_3^5$ | $C_4^5$ | $C_5^5$ |
| $P_1$ | 9 | 210 | 32 | 5 | 193 | 28 | 32 | 3 | 0 |
| $P_2$ | 24 | 0 | 159 | 24 | 0 | 76 | 131 | 0 | 0 |
| $P_3$ | 76 | 0 | 0 | 15 | 0 | 0 | 1 | 79 | 11 |
| $P_4$ | 0 | 0 | 3 | 76 | 0 | 4 | 5 | 0 | 70 |

Table 2 shows a detailed comparison between PAM and CoRe results: when $K = 4$ both algorithms seem to agree on the existence of a large $C_2^4$-$P_1$ group, that CoRe characterizes with an high PR-relevance and a medium relevance of the ER marker (see Fig.3.a). Part of PAM's $P_1$ cluster is split by CoRe and assigned to $C_3^4$ (roughly corresponding to PAM's $P_2$), which has a very similar pattern of feature relevance with respect to $C_2^4$, except for a slightly higher significance of the ER covariate. The two smaller groups identified by PAM, i.e. $P_3$ and $P_4$, roughly correspond to CoRe's $C_1^4$ and $C_4^4$ groups, respectively. CoRe clusters, however, are larger than PAM's $P_3$ and $P_4$, since they gather samples from the highly populated $P_2$ group. Interestingly, CoRe isolates sharply the small $P_3$ and $P_4$ groups in the $K = 5$ hypothesis: the results in Table 2 show a strong $P_3$-$C_4^5$ and $P_4$-$C_5^5$ overlap. Seemingly, the addition of the cluster $C_2^5$ attracted those *spurious* $P_2$ samples that, in the $K = 4$ hypothesis, were assigned $C_1^4$ and $C_4^4$. In our opinion, this suggests the existence of a fifth cluster that is situated *in between* the $P_2$, $P_3$ and $P_4$ groups from PAM.

The results in [3] showed that the $P_2$ samples are associated with less aggressive tumor features, while individuals from the $P_3$ and $P_4$ groups tend to develop an increased number of metastatic lymph nodes. Therefore, the additional cluster discovered by CoRe can possibly describe a bioprofile characterizing patients that may or may not develop an aggressive form of breast cancer, or develop metastatic lymph nodes at a latter stage of the tumor development.

## 4 Conclusion

We have studied unsupervised cluster number estimation on a breast cancer dataset, comparing the performance of hard clustering, neural and model based algorithms. The experimental results showed how CoRe can be used to unsupervisedly explore biomedical data, estimating the number of the clusters in the dataset and producing a measure of the relevance of the samples' covariates that can be used to identify significant markers in the discovered bio-profiles. The results of the Bayesian models suggests that their performance could have been affected by the nature of the dataset, preventing them from agreeing on a common data partition and a stable cluster number estimate. The results produced by CoRe seem to confirm the hypothesis presented in [3] concerning the existence of a fifth tumor subgroup in the breast cancer case-study. Our hypothesis, that shall by validated by clinical studies, is that such sub-group might differentiate a form of breast cancer characterized by an initial, apparent, low aggressiveness that might later evolve in a more aggressive, metastatic, form.

## Acknowledgements

This research was performed with the financial support of the Italian Ministry of University and Research (MIUR), PRIN prot. 2005030003-005. This work was carried out with partial funding from the Biopattern Network of Excellence FP6/2002/IST/1; N IST-2002-508803 (http://www.biopattern.org).

## References

1. Bacciu, D., Starita, A.: A robust bio-inspired clustering algorithm for the automatic determination of unknown cluster number. In: International Joint Conference on Neural Networks (IJCNN 2007), pp. 1314–1319. IEEE, Los Alamitos (2007)
2. Bacciu, D., Micheli, A., Starita, A.: Simultaneous clustering and feature ranking by competitive repetition suppression learning with application to gene data analysis. In: Computational Intelligence in Medicine and Healthcare (CIMED 2007) (2007)
3. Ambrogi, F., Biganzoli, E., Querzoli, P., Ferretti, S., Boracchi, P., Alberti, S., Marubini, E., Nenci, I.: Molecular subtyping of breast cancer from traditional tumor marker profiles using parallel clustering methods. Clin. Cancer Res. 12(1), 781–790 (2006)
4. Kohonen, T.: Self-organized formation of topologically correct feature maps. Biol. Cybern. 43(1), 59–69 (1982)
5. Carpenter, G., Grossberg, S.: The ART of adaptive pattern recognition by a self-organizing neural network. Computer 21(3), 77–88 (1988)
6. Figueiredo, M.A.T., Jain, A.K.: Unsupervised learning of finite mixture models. IEEE Trans. Pattern Anal. Mach. Intell. 24(3), 381–396 (2002)
7. Corduneanu, A., Bishop, C.: Variational bayesian model selection for mixture distributions. In: Richardson, T., Jaakkola, T. (eds.) Artificial Intelligence and Statistics, pp. 27–34. Morgan Kaufmann, San Francisco (2001)
8. Constantinopoulos, C., Likas, A.: Unsupervised learning of gaussian mixtures based on variational component splitting. IEEE Trans. Neural Netw. 18(3), 745–755 (2007)

# Exploratory Characterization of Outliers in a Multi-centre $^{1}$H-MRS Brain Tumour Dataset

Alfredo Vellido[1,⋆], Margarida Julià-Sapé[2,3], Enrique Romero[1],
and Carles Arús[3,2]

[1] Dept. de Llenguatges i Sistemes Informàtics - Universitat Politècnica de Catalunya
C. Jordi Girona, 1-3. 08034, Barcelona, Spain
{avellido,eromero}@lsi.upc.edu
http://www.lsi.upc.edu/~websoco/AIDTumour

[2] Centro de Investigación Biomédica en Red en Bioingeniería, Biomateriales y
Nanomedicina (CIBER-BBN), Cerdanyola del Vallès, Spain

[3] Grup d'Aplicacions Biomèdiques de la RMN (GABRMN)
Departament de Bioquímica i Biología Molecular (BBM). Unitat de Biociències
Universitat Autònoma de Barcelona (UAB), Cerdanyola del Vallès, Spain
marga@carbon.uab.es,carles.arus@uab.es

**Abstract.** As part of the AIDTumour research project, the analysis of MRS data corresponding to various tumour pathologies is used to assist expert diagnosis. The high dimensionality of the MR spectra might obscure atypical aspects of the data that would jeopardize their automated classification and, as a result, the process of computer-based diagnostic assistance. In this paper, we put forward a method to overcome this potential problem that combines automatic outlier detection, visualization through dimensionality reduction, and expert opinion.

**Keywords:** Proton Magnetic Resonance Spectroscopy, Brain Tumours, Outlier Detection, Data exploration, Data Visualization, Dimensionality Reduction; Medical Decision Support Systems.

## 1 Introduction

Decision making in oncology is a sensitive matter, and even more so in the specific area of brain tumour oncologic diagnosis, for which the direct and indirect costs - both human and financial - of misdiagnosis are very high. In this area, in which most diagnostic techniques must be non-invasive, clinicians should benefit from the use of an at least partially automated computer-based medical Decision Support System (DSS).

AIDTumour (Artificial Intelligence Decision Tools for Tumour diagnosis [1]) is a research project for the design and implementation of a medical DSS to assist experts in the diagnosis of human brain tumours on the basis of data

[⋆] A. Vellido is a Spanish M.E.C. Ramón y Cajal researcher. The authors acknowledge funding from M.E.C. projects TIN2006-08114 and SAF2005-03650, and, from 1$^{st}$ January 2003, Generalitat de Catalunya (grant CIRIT SGR2005-00863).

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 189–196, 2008.

obtained by Magnetic Resonance Spectroscopy (MRS). This is a technique that can shed light on cases that remain ambiguous after clinical investigation. The MRS data used in AIDTumour and analyzed in this paper belong to a complex multi-centre set containing cases of several brain tumour pathologies [2]. These data have undergone a rigorous pre-processing quality control that validates them from the viewpoint of the radiologists. Nevertheless, and for their use in an automated computer-based DSS, the various origins of these spectra and the complexity of their pre-processing make further data exploration advisable.

It might be problematic to include some of the spectra in an automated DSS without further ado for three different reasons: Firstly, some may contain measurement or acquisition artifacts that, even if not completely precluding diagnosis by visual inspection, might induce errors in computer-based diagnosis: these are what we call here *artifact-related outliers*. Secondly, atypical cases that do not contain artifacts but are nevertheless unrepresentative of the main distributions of the whole dataset: herein, these will be referred to as *distinct outliers* [3]. Thirdly, some cases with a clear biopsy-based diagnosis (tumour type attribution) may yield spectra that are quantitatively similar to those of other tumour types, misleading a computer-based classification system. Even if representative of the data as a whole, they are still unrepresentative of their own tumour type: these we will call *class outliers*. Note that these three reasons are not always mutually exclusive.

In this paper, we show the effectiveness of a method to identify and characterize potentially conflicting MRS data that combines techniques of dimensionality reduction, exploratory visualization, and outlier detection, with expert knowledge. The introduction of the latter is paramount, as it will help to skim those cases truly conflictive out of those shortlisted by blind quantitative criteria. Overall, this method is conceived as a preliminary step to data classification in the DSS. Dimensionality reduction is not trivial in this setting, as the available MRS data are scarce and high dimensional. Sammon's mapping [4] is used to this end. Generative Topographic Mapping (GTM [5]), a manifold learning model, is used to quantify spectra atypicality.

## 2 MRS Data

The analysed MRS data correspond to 217 short-echo time (SET) and 195 long-echo time (LET) single voxel $^1$H MR spectra acquired in vivo from brain tumour patients. They include 58 (SET) and 55 (LET) meningiomas (*mm*), 86 (SET) and 78 (LET) glioblastomas (*gl*), 38 (SET) and 31 (LET) metastases (*me*), 22 (SET) and 20 (LET) astrocytomas grade II (*a2*), 6 (SET and LET) oligoastrocytomas grade II (*oa*), and 7 (SET) and 5 (LET) oligodendrogliomas grade II (*od*). For details on data acquisition and processing, see [2]. Class labelling was performed according to the World Health Organization (WHO) system for diagnosing brain tumours by histopathological analysis of a biopsy sample. For the reported analysis, spectra were bundled into three groups, namely: G1: *low grade gliomas* (*a2*, *oa* and *od*); G2: *high grade malignant tumours* (*me* and *gl*); and

G3: *meningiomas*. The clinically-relevant regions of the spectra were sampled to obtain 195 frequency intensity values (measured in parts per million (ppm), an adimensional unit of relative frequency position in the data vector), from 4.25 parts per million (ppm) down to 0.56 ppm, which become data attributes.

# 3 Methods

## 3.1 MRS Data Visualization through Sammon's Mapping

In order to allow the visualization of the data through dimensionality reduction, the spectra were mapped onto a 3-D space through Sammon's mapping [4]. The non-linear mapping is constructed as to minimize the inter-point distortions it introduces, quantified by Sammon's error measure:

$$\frac{1}{\sum_{i<j} \delta_{ij}} \sum_{i<j} \frac{(\delta_{ij} - \xi_{ij})^2}{\delta_{ij}}, \tag{1}$$

where $\delta_{ij}$ is the Euclidean distance between spectra $i$ and $j$ in the original data space and $\xi_{ij}$ is the Euclidean distance between the projections of these spectra in the 3-D space. In this study, the minimization of the Sammon's error was performed by the Newton method. A collection of models was obtained by varying the initial points (100 different random values) and the step size (9 different values), for a total of 900 runs. The models with lowest Sammon's error were selected for further analysis.

## 3.2 Outlier Detection Using *t*-GTM

Generative Topographic Mapping (GTM [5]) is a non-linear latent variable model defined as a mapping from a low dimensional latent space onto the multivariate data space. The mapping is carried through by a set of basis functions and is defined as a generalized linear regression model:

$$\mathbf{y} = \phi(\mathbf{u})\mathbf{W}, \tag{2}$$

where $\mathbf{W}$ is a matrix of adaptive weights that defines the mapping, and $\mathbf{u}$ is a point in latent space. $\phi$ are $M$ basis functions that, in the original formulation, were chosen to be spherically symmetric Gaussians. For this Gaussian GTM, the presence of outliers is likely to negatively bias the estimation of its adaptive parameters. In order to overcome this limitation, the GTM was recently redefined [3] as a constrained mixture of Student's $t$ distributions: the $t$-GTM. The mapping described by Equation (2) remains, with the basis functions now being Student's $t$ distributions. As a byproduct of this reformulation of GTM, and following [7], a statistic quantifying to what extent $t$-GTM considers a data case to be an outlier can be defined as $O_n = \sum_k p(\mathbf{u}_k|\mathbf{x}_n)\beta\|\mathbf{y}_k - \mathbf{x}_n\|^2$, where $\beta$ is the inverse of the noise variance. The larger the value of this statistic the more likely the case is to be an outlier. Notice that $p(\mathbf{u}_k|\mathbf{x}_n)$ is the responsibility assumed by a latent point $k : 1, \ldots, K$ for the data case $n$ and, the same as for the standard GTM, it is obtained as part of the maximum likelihood estimation of the model's parameters.

### 3.3 Shortlisting Outlier Cases of Interest

The free software package KING [6] is used to visualize in 3-D the Sammon's mapping of the spectra described in section 2, enabling a preliminary data exploration. The data projections obtained with Sammon's mapping were then modelled by $t$-GTM, obtaining a value of $O_n$ for each data case, indicating the corresponding degree of atypicality. Histograms of $O_n$ were generated to shortlist potentially conflictive cases of the three types described in the introduction. Loose thresholds of the statistic were set for the selection of the lists of outlier candidates. Using all this information, an expert in MRS then singled out those spectra she/he considered to be truly atypical in any sense and compared them to the characteristic spectra corresponding to their tumour type.

## 4 Experimental Results and Discussion

### 4.1 Short Echo Time $^1$H MRS Data

The histogram in Fig. 1 displays the distribution of the value of the statistic $O_n$, first calculated for the complete SET MRS dataset. A threshold of $O_n = 20$ was set to shortlist outlier candidate spectra. This yielded 23 potential outliers, which were inspected by an expert who decided that only 19 of them (4 *distinct outliers* and 15 *artifact-related outliers*) qualified as such, for different causes listed in Table 1 (left). Notice that there are plenty of *low grade gliomas* (37% of all outliers, while only 16% of all data). Six different types of artifacts were found in the data, namely: spectra heavily contaminated by noise; bad water signal suppression as part of the data pre-processing; incorrect spectrum alignment of the ppm reference; incorrect baseline; the existence of polispiculated artifact; and signal distorsion due to eddy currents (induced as a result of field gradient switching in signal acquisition).

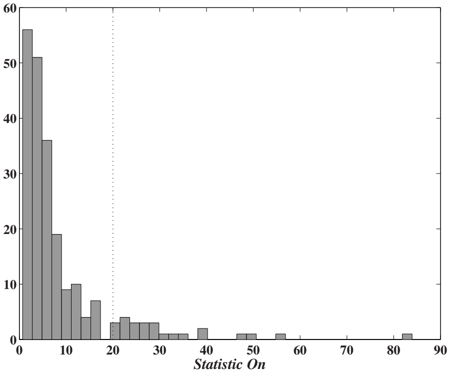


**Fig. 1.** Histogram of statistic $O_n$ for the SET dataset. The selected threshold at value 20 is represented as a vertical dotted line.

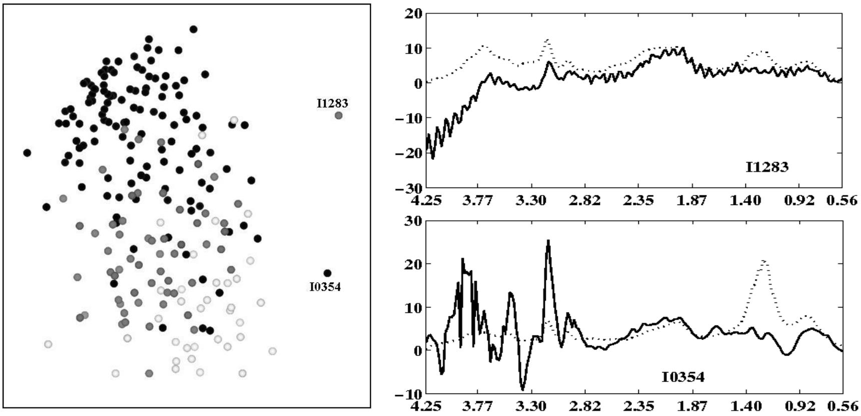


**Fig. 2.** 3-D Sammon's mapping view of two cases of interest (with groups of tumours displayed in different shades of gray), on the left column, and their corresponding individual spectra (solid lines) and mean spectra (dotted lines) of the tumour groups they belong to, on the right column. The abscissa axis displays frequency in ppm.

To illustrate the visualization of the high-dimensional spectra through Sammon's mapping, Fig. 2 displays SET cases I1283 (a meningioma, which the expert described as being contaminated by noise, and affected by bad water suppression, polispiculated effect and eddy currents), and I0354 (a glioblastoma, which the expert described as being affected by a polispiculated effect). Their atypicality is clearly captured by the visualization.

Spectra can also be atypical specifically with respect to their group of tumours. These are what we call *class outliers*. The histograms of $O_n$ for each group of tumours are omitted here for the sake of brevity. Five *low grade gliomas*, 20 *high grade malignant tumours*, and 13 *meningiomas* where shortlisted and inspected by the expert, who considered that, out of these, none of the *low grade gliomas*, only 9 *high grade malignant tumours*, and 8 *meningiomas* should be tagged as *class outliers*. Some of them also contain artifacts, given that, as mentioned in the introduction, *artefact-related outliers* and *class outliers* are not mutually exclusive characterizations. They are described in Table 2 (left). It is very interesting that, even though *low grade glioma* outliers are plentiful, as seen in Table 1 (left), there is no *class outlier* amongst them in the SET spectra, suggesting a well-defined against the rest but less-than-compact structure in this group of tumours.

### 4.2 Long Echo Time $^1$H MRS Data

The histogram in Fig. 3 displays the distribution of $O_n$ for the complete LET MRS dataset. A threshold of $O_n = 15$ was set to shortlist outlier candidate spectra. This yielded 21 potential outliers, which were again inspected by an

**Table 1.** Outlier characterization of the SET (left) and LET (right) $^1$H MRS datasets. Columnwise, *Id* is an anonymized case identifier; star superscripts indicate that there are artifacts that do not preclude the expert's correct interpretation of the case. *Tum* refers to tumour type (see labels in section 2). *Dis* refers to *Distinct outliers*. Six types of artifacts were found: *noi* stands for noise; *wat*, for bad water signal suppression; *ali*, for alignment; *lin*, linebase; *pol*, for the polispiculated effect; and *edd* for eddy currents. See main text for details.

| ***Id*** | ***Tum*** | ***Dis*** | ***Artifact-relat. outl.*** | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | noi | wat | ali | bas | pol | edd |
| I0335 | G1(a2) | X | | | | | | |
| I1052$^*$ | G1(a2) | | X | | | | | |
| I1087$^*$ | G1(a2) | | | X | | | | |
| I0060 | G1(oa) | | X | | | | | |
| I0069 | G1(oa) | | | | X | | | |
| I0450 | G1(oa) | X | | | | | | |
| I0179 | G1(od) | X | | | | | | |
| I0135$^*$ | G2(gl) | | | | | X | | |
| I0172$^*$ | G2(gl) | | | X | | X | | |
| I0354$^*$ | G2(gl) | | | | | | X | |
| I0421$^*$ | G2(gl) | | | X | | | | |
| I1024$^*$ | G2(gl) | | | X | | | | |
| I0055 | G2(me) | | X | | | | | |
| I0244$^*$ | G3(mm) | X | | | | | | |
| I0375 | G3(mm) | | X | | | | | |
| I0381$^*$ | G3(mm) | | | X | | | | |
| I0390$^*$ | G3(mm) | | | | | | X | |
| I0393$^*$ | G3(mm) | | X | | | X | X | |
| I1283$^*$ | G3(mm) | | X | X | | | X | X |

| ***Id*** | ***Tum*** | ***Dis*** | ***Artifact-relat. outl.*** | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | noi | wat | ali | bas | pol | edd |
| I1061 | G1(a2) | | | | X | | | |
| I0062$^*$ | G2(gl) | | X | X | | X | | |
| I0105$^*$ | G2(gl) | X | | | | | | |
| I0172 | G2(gl) | | | X | | X | | |
| I0175$^*$ | G2(gl) | | X | | | | X | |
| I0354$^*$ | G2(gl) | | | X | | | X | |
| I0428$^*$ | G2(gl) | | | X | | | X | |
| I1044$^*$ | G2(gl) | | | | | | X | |
| I1057$^*$ | G2(gl) | | X | X | | | X | |
| I1379$^*$ | G2(gl) | | X | | | | | X |
| I0027 | G2(me) | | X | | X | | | |
| I0368$^*$ | G2(me) | | | X | | | X | |
| I1070 | G2(me) | X | | | | | | |
| I0390$^*$ | G3(mm) | | | | | | X | |
| I0420 | G3(mm) | | | | | | X | |
| I1074 | G3(mm) | | X | | | | | |
| I1090 | G3(mm) | X | | | | | | |
| I1378 | G3(mm) | | X | | | | X | |

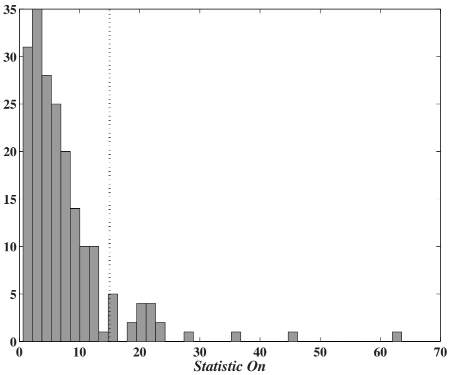


**Fig. 3.** Histogram of statistic $O_n$ for the LET dataset. The selected threshold at value 15 is represented as a vertical dotted line.

expert, who decided that only 18 of them qualified as such (3 *distinct outliers* and 15 *artifact-related outliers*). The corresponding characterization is presented in Table 1 (right). Interestingly, in this case there is almost no *low grade glioma* outlier and, instead, *high grade malignant* outliers predominate (67% of all outliers, while only 56% of all data).

Turning now our attention to *class outliers*, 9 *low grade gliomas*, 7 *high grade malignant tumours*, and 10 *meningiomas* were shortlisted and inspected by the expert, who considered that, out of these, none of the *low grade gliomas*, only 2 *high grade malignant tumours*, and 5 *meningiomas* should be tagged as *class outliers*. Some of them also contain artifacts, and they are characterised in Table 2 (right). It is worth noting that there are far less *class outliers* in the LET dataset than in the SET one, suggesting a much more compact definition of the tumour groups in the former representation. It is also interesting that, again, there is no *class outlier* amongst the *low grade gliomas*. Together with the almost complete lack of outliers in this tumour group shown in Table 1 (right), this indicates that they have a much more compact and well-defined structure in the LET representation.

**Table 2.** *Class outlier* characterization of the SET (left) and LET (right) $^1$H MRS datasets, by groups of tumours. Label description as in Table 1.

| ***Id*** | ***Tum*** | ***Artifacts*** | | | | | |
|---|---|---|---|---|---|---|---|
| | | noi | wat | ali | bas | pol | edd |
| ***Low grade gliomas*** (G1) | | | | | | | |
| ∅ | | | | | | | |
| ***High grade malignant*** (G2) | | | | | | | |
| I0021* | gl | | | | | | |
| I0358* | gl | | | | | X | |
| I0200* | gl | | | | | X | |
| I1390 | gl | | | X | | | |
| I0168* | gl | | | | | | |
| I1098* | gl | | | | | | |
| I1076* | me | | | | X | | |
| I0352* | me | | | | X | | |
| I1377* | me | | | | | | |
| ***Meningiomas*** (G3) | | | | | | | |
| I0160* | mm | | | | | | |
| I1090* | mm | | | | | | |
| I1073* | mm | | | | | X | |
| I0009 | mm | | | | | | |
| I0390* | mm | | | | | X | |
| I1378* | mm | | | | X | | |
| I0375 | mm | X | | | | X | |
| I1149* | mm | | | | | | |

| ***Id*** | ***Tum*** | ***Artifacts*** | | | | | |
|---|---|---|---|---|---|---|---|
| | | noi | wat | ali | bas | pol | edd |
| ***Low grade gliomas*** (G1) | | | | | | | |
| ∅ | | | | | | | |
| ***High grade malignant*** (G2) | | | | | | | |
| I0105* | gl | | | | | | |
| I1070 | me | | | | | | |
| ***Meningiomas*** (G3) | | | | | | | |
| I0114* | mm | | | | | | |
| I1090 | mm | | | | | | |
| I1378 | mm | X | | | | X | |
| I0002* | mm | | | | | | |
| I0009* | mm | | | | | | |

## 5 Conclusion

In this paper, we have defined a method to identify and characterize potentially conflicting MRS multi-center data corresponding to several brain tumour pathologies, which combines dimensionality reduction, outlier detection and exploratory visualization techniques with expert knowledge. This combination of data-based analysis and human expertise is one of the distinctive hallmarks of Evidence-Based Medicine (EBM) for healthcare practice [8]. This method will be embedded in a medical DSS resulting from the AIDTumour [1] research project.

Several research questions would require further research. First, the usefulness of outlier detection and characterization for the improvement of automated tumour diagnostic classification should be assessed. For instance, does the fact the LET MRS data include less *class outliers* mean that we should expect better classification of tumour groups using these and not SET data? Second, only 2 glioblastomas and 1 meningioma tagged as *artifact-related outliers*, and 3 meningiomas tagged as *class outliers* appear both for SET and LET data. How can we explain this level of mismatch? Specific policies to deal with this wide variety of situations should be carefully implemented in the projected DSS.

**Acknowledgments.** Authors gratefully acknowledge the former INTERPRET (EU-IST-1999-10310) European project partners. Data providers: Dr. C. Majós (IDI), Dr.À. Moreno-Torres (CDP), Dr. F.A. Howe and Prof. J. Griffiths (SGUL), Prof. A. Heerschap (RU), Dr. W. Gajewicz (MUL) and Dr. J. Calvar (FLENI); data curators: Dr. A.P. Candiota, Ms. T. Delgado, Ms. J. Martín, Mr. I. Olier and Mr. A. Pérez (all from GABRMN-UAB). C. Arús and M. Julià-Sapé are funded by the CIBER of Bioengineering, Biomaterials and Nanomedicine, an initiative of the *Instituto de Salud Carlos* III (ISCIII) of Spain.

## References

1. Artificial Intelligence Decision Tools for Tumour diagnosis research project, `http://www.lsi.upc.edu/~websoco/AIDTumour`
2. Julià-Sapé, M., et al.: A Multi-Centre, Web-Accessible and Quality Control-Checked Database of in Vivo MR Spectra of Brain Tumour Patients. Magn. Reson. Mater. Phy. 19, 22–33 (2006)
3. Vellido, A., Lisboa, P.J.G.: Handling Outliers in Brain Tumour MRS Data Analysis through Robust Topographic Mapping. Comput. Biol. Med. 36, 1049–1063 (2006)
4. Sammon Jr., J.W.: A nonlinear mapping for data structure analysis. IEEE T. Comput. C-18, 401–409 (1969)
5. Bishop, C.M., Svensén, M., Williams, C.K.I.: The Generative Topographic Mapping. Neural Comput. 10(1), 215–234 (1998)
6. KING visualization software, `http://kinemage.biochem.duke.edu/software/king.php`
7. Peel, D., McLachlan, G.J.: Robust mixture modelling using the t distribution. Stat. Comput. 10, 339–348 (2000)
8. Dickersin, K., Straus, S.E., Bero, L.A.: Evidence Based Medicine: Increasing, not Dictating. Choice. Brit. Med. J. 334(suppl.1), s10 (2007)

# Feature Selection in *in vivo* $^1$H-MRS Single Voxel Spectra

Félix F. González-Navarro and Lluís A. Belanche-Muñoz

Dept. de Llenguatges i Sistemes Informàtics
Universitat Politècnica de Catalunya
*Ω*-Building, North Campus. Barcelona, Spain
{fgonzalez,belanche}@lsi.upc.edu

**Abstract.** Machine learning is a powerful paradigm within which to analyze $^1$H-MRS spectral data for the classification of tumour pathologies. An important characteristic of this task is the high dimensionality of the involved data sets. In this study we apply several feature selection algorithms in order to reduce the complexity of the problem on two types of $^1$H-MRS spectral data: long-echo and short-echo time, which present considerable differences in the spectrum for the same cases. The obtained experimental results show that the feature selection algorithms enhance the classification performance of the final induced models both in terms of prediction accuracy and number of involved spectral frequencies. The results obtained using a fast algorithm based on entropic measures of subsets of spectral data are specially promising.

## 1 Introduction

*In vivo* nuclear proton magnetic resonance spectroscopy ($^1$H-MRS) is a powerful technique that helps to observe metabolic processes in living tissue [1]. Although these metabolic functions are not fully understood, it is possible to employ *machine learning* techniques for the diagnosis and grading of adult brain tumours [2]. Several recent examples in the literature use machine learning techniques for distinguishing between different brain tumour pathologies (e.g. [3], [4]). Due to the high dimensionality (near 200 spectral measurements in the present study), these efforts use dimensionality reduction methods (feature selection and/or extraction) to lower the complexity of the problem. Of added practical importance is the interpretability of the solutions in terms of the obtained spectral frequencies, which limits the applicability of methods such as PCA or ICA.

In a previous work, $^1$H-MRS long-echo data were analysed with the purpose of obtaining classification models showing good generalization ability after a strong dimensionality reduction process [5]. In the present study we are interested in performing a more detailed feature selection study in both long- and short-echo types of data, treating long-echo and short-echo time spectral points as the features. In particular, we use an Entropic Filtering Algorithm (EFA) for feature selection as a fast method to generate a relevant subset of spectral frequencies. Performance of other fast feature selection algorithms is also reported.

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 197–205, 2008.

The first goal is to obtain simple models (in terms of low numbers of features) that generalize well. The second goal is to progress in the direction of assessing the differential performance for the two types of spectra, which present notable differences for the same cases.

## 2 An Entropic Filtering Algorithm

Mutual Information (MI) measures the mutual dependence of two random variables. It has been used with success as a criterion for feature selection in machine learning tasks. In this work we use this concept embedded in a fast algorithm that computes MI between a set of variables and the class variable by generating first a “super-feature”, obtained considering the concatenation of each combination of possible values of its forming features. In symbols, let $X = \{X_1, ..., X_n\}$ be the original feature set and consider a subset $\tau = \{\tau_1, \cdots, \tau_k\}$. A single feature $\mathcal{V}_\tau$ can be obtained uniquely, whose possible values are the concatenations of all possible values of the features in $\tau$. The conditional entropy between $\mathcal{V}\tau$ and the class feature $Y$ is then:

$$H(Y|\tau_1, \cdots, \tau_k) = H(Y|\mathcal{V}_\tau) = -\sum_{v \in \mathcal{V}_\tau} \sum_{y \in Y} p(v, y)\, log \frac{p(v, y)}{p(y)}. \tag{1}$$

Proceeding in this way, the MI can be determined as a simple bivariate case: $I(\mathcal{V}_\tau; Y) = H(Y) - H(Y|\mathcal{V}_\tau)$. An *index of relevance* of the feature $X_i \in X$ to a class $Y$ with respect to a subset $\tau \subset X$, inspired on [6], is given by:

$$R(X_i; Y|\tau) = \frac{I(X_i; Y|\mathcal{V}_\tau)}{H(Y|\mathcal{V}_\tau)} = \frac{H(Y|\mathcal{V}_\tau) - H(Y|X_i; \mathcal{V}_\tau)}{H(Y|\mathcal{V}_\tau)}. \tag{2}$$

This index of relevance of a feature subset is to be maximized (it has a maximum value of 1). This measure is used in this study to evaluate subsets of spectral frequencies, embedded into a filter forward-selection strategy, conforming the *Entropic Filtering Algorithm* (EFA). A detailed description and a fast implementation of the whole algorithm can be found in [5].

In order to apply the algorithm, a discretization process is needed. Many dimensionality reduction studies use discretization schemes as a way to favor classification tasks (such as [7], [8]). This change of representation does not often result in a significant loss of accuracy (sometimes significantly improves it); it also offers large reductions in learning time. The CAIM algorithm [9] is selected because it is able to work with supervised data and does not require the user to define a specific number of intervals for each feature.

## 3 Experimental Work

The echo time is an influential parameter in $^1$H-MRS spectra acquisition. In short-echo time spectra (typically 20-40 ms) some metabolites are better evaluated (e.g. lipids, myo-inositol, glutamine and glutamate). However, there may

be numerous overlapping resonances (e.g. glutamate/glutamine at 2.2 ppm and NAA at 2.01 ppm) which make the spectra difficult to interpret [10]. A long-echo time (270-288 ms) yields less metabolites but also less baseline distortion, resulting in a more readable spectrum. There are a few studies comparing the classification potential of the two types of spectra (see e.g. [10], [11]). These works seem to give a slight advantage to using short-echo time information or else suggest a combination of both types of spectra.
The analyzed $^1$H-MRS dataset is detailed as follows:

- 266 single voxel *long-echo time* spectra acquired *in vivo* from brain tumour patients, out of which 195 are used in this study, including: meningiomas (55 cases), glioblastomas (78), metastases (31), astrocytomas Grade II (20), oligoastrocytomas Grade II (6) and oligodendrogliomas Grade II (5);
- 304 single voxel *short-echo time* spectra, of which 217 are used: meningiomas (58 cases), glioblastomas (86), metastases (38), astrocytomas Grade II (22), oligoastrocytomas Grade II (6) and oligodendrogliomas Grade II (7).

Class labelling was performed according to the World Health Organization system for diagnosing brain tumours by histopathological analysis of a biopsy sample. Both spectra were grouped into three superclasses: high-grade malignant tumours (metastases and glioblastomas), low-grade gliomas (astrocytomas, oligodendrogliomas and oligoastrocytomas) and meningiomas. The spectra consist of 195 frequency intensity values, from 4.21 ppm down to 0.51 ppm.

### 3.1 Experimental Setup

The $^1$H-MRS data sets were randomly split into two parts: 70% for the feature selection process itself, *a posteriori* classifier induction and model selection (hereafter called the *training* set), and the remaining 30%, used to ascertain the generalization ability of the classifiers (the *test* set). This division was done keeping the relative proportion of classes in the whole data, yielding 53 and 67 test observations for long-echo and short-echo time data, respectively.

Four different classifiers were designed using the training set by means of 3-fold cross validation and the full set of frequencies. The classifiers are: the nearest-neighbour technique with Euclidean metric (*NN*) and parameter $k$ (number of neighbours), the *Naïve Bayes classifier* (NB), a *C4.5* decision tree with parameter *cp* (complexity parameter) and a Random Forest (RF) [12] with parameter *nt* (number of trees). The EFA is applied to the discretized $^1$H-MRS data (the training part) to obtain what will be called Best Spectral Subset (BSS). Note that the EFA does not need an inducer. The four classifiers can then be built in the training sets using the continuous frequencies (both in the full set and in the obtained BSSs) and evaluated in the corresponding test sets.

The filtering algorithm Relief [13] is also used as a feature selection method for comparative purposes. This is actually a feature-weighting algorithm that takes feature interactions into account and yields a set of feature weights that can be sorted in descending order. A cut-point was established according to the Pareto principle to obtain a feature subset. For many events, 80% of the effects (viz. classification ability) come from 20% of the causes (viz. spectral frequencies). To this end, the obtained weights are linearly transformed so that the smallest weight equals zero and their sum equals one. Then the top ranking features such that their accumulated weight is closest to 0.2 are selected.

Forward Selection in wrapper mode [14] is also used in the comparison. This algorithm incrementally adds one feature at each step: the one maximizing a given criterion function. In this study, the classifiers referenced above were used as the criterion function by means of 3-fold cross-validation in the training set. In the next section, for every feature selection experiment, the size of the corresponding BSSs, their accuracy (Acc) and macro-averaged $F_1$-measure[1] on the test set are reported, as well as the parameter values found by model selection. This is done separately for the long-echo and short echo $^1$H-MRS data sets.

### 3.2 Long-Echo Time Experimental Results and Discussion

Long-echo time BSSs are reported in table 1 (left part). Five subsets (named BSS-L(1) to BSS-L(5)) reached maximum relevance ($R = 1$), with just seven spectral frequencies. The results for the four classifiers using the full set of frequencies (indicated with NR), the corresponding results using BSS-L(1) to BSS-L(5) and the Relief solution are displayed in table 2. Finally, the results using Forward Selection in wrapper mode are presented in table 3 (left).

**Table 1.** Five spectral subsets found by the EFA that reach maximum relevance ($R = 1$) on long-echo time $^1$H-MRS data (left table) and two on short-echo time (right table)

| BSS-L | Spectral frequencies (ppm) |
|---|---|
| (1) | 3.22 3.03 2.54 2.48 2.16 1.57 1.27 |
| (2) | 3.03 2.79 2.54 2.48 2.16 1.57 1.27 |
| (3) | 3.03 2.54 2.48 2.20 2.16 1.57 1.27 |
| (4) | 3.03 2.54 2.48 2.16 1.57 1.47 1.27 |
| (5) | 3.03 2.54 2.48 2.16 1.57 1.27 1.21 |

| BSS-S | Spectral frequencies (ppm) |
|---|---|
| (1) | 3.70 3.03 2.37 2.10 1.45 1.04 |
| (2) | 3.70 3.50 3.03 2.37 1.45 1.04 |

The best results are obtained by evaluating NN on the EFA solutions BSS-L(1) and BSS-L(5). The positions in the spectrum of these two subsets of frequencies (of size seven) are depicted in Fig. 1, shown against average spectra per class. Some of the selected frequencies can be related to known metabolites: 3.22 corresponds to the Choline peak; 3.03 to the Creatine peak; 2.16, 2.48 and 2.54 are roughly in the area of glutamine-glutamate and lipid/macromolecule peaks; 1.57 is located nearby the Alanine peak; finally, 1.21 and 1.27 are within the lipids peak area. Using either of these subsets, test set performance is worse (for C4.5 and NB), slightly better (for RF) or much better (for NN), compared to using the full set of frequencies, providing evidence in favor of a feature selection

[1] This is the arithmetic average of the $F_1$-measures (harmonic mean of precision and recall) obtained for the three tumor super-classes. Macro-averaged $F_1$-measure gives equal weight to each class and thus is more influenced by performance on rare classes.

**Table 2.** Test set performance on the $^{1}$H-MRS long-echo time. NR = no reduction, $k$ = number of neighbours, $cp$ = complexity parameter, $nt$ = number of trees. Best results have been highlighted in bold.

| Reduction | BSS | NN | | | NB | | C4.5 | | | RF | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| method | size | Acc | $F_1$ | $k$ | Acc | $F_1$ | Acc | $F_1$ | $cp$ | Acc | $F_1$ | $nt$ |
| NR | 195 | 85.0% | 81.0% | 9 | 90.5% | 88.1% | 83.0% | 77.2% | 0.25 | 83.0% | 75.4% | 9 |
| BSS-L(1) | 7 | **94.3%** | **92.5%** | 12 | 88.6% | 86.0% | 75.5% | 71.0% | 0.25 | 83.0% | 80.6% | 5 |
| BSS-L(2) | 7 | 92.4% | 91.3% | 8 | 85.0% | 81.5% | 75.5% | 70.0% | 0.25 | 84.9% | 82.6% | 15 |
| BSS-L(3) | 7 | 90.5% | 82.2% | 5 | 85.0% | 82.5% | 75.4% | 69.7% | 0.25 | 88.6% | 86.0% | 15 |
| BSS-L(4) | 7 | 90.5% | 87.8% | 15 | 86.8% | 84.2% | 75.4% | 69.8% | 0.25 | 84.9% | 84.4% | 14 |
| BSS-L(5) | 7 | **94.3%** | **93.1%** | 13 | 86.8% | 85.1% | 79.2% | 73.2% | 0.25 | 85.0% | 82.6% | 8 |
| Relief | 15 | 71.7% | 61.0% | 10 | 60.4% | 52.7% | 62.2% | 54.6% | 0.25 | 71.7% | 64.4% | 5 |

**Table 3.** Left: $^{1}$H-MRS long-echo time test set results using Forward Subset Selection (FSS) in wrapper mode. Center and right: test set confusion matrices using NN on the subsets BSS-L(1) and BSS-L(5), respectively. True class falls vertically.

| method | size | Acc | $F_1$ | $k/cp$ |
|---|---|---|---|---|
| FSS-NN | 12 | 88.7% | 87.0% | 4 |
| FSS-NB | 11 | 81.1% | 78.0% | - |
| FSS-C4.5 | 7 | 73.6% | 67.0% | 0.25 |

| True | **BSS-L(1)+NN** | | |
|---|---|---|---|
| Class | **LG** | **HG** | **ME** |
| **LG** | 6 | 2 | 0 |
| **HG** | 0 | 30 | 0 |
| **ME** | 0 | 1 | 14 |

| True | **BSS-L(5)+NN** | | |
|---|---|---|---|
| Class | **LG** | **HG** | **ME** |
| **LG** | 7 | 1 | 0 |
| **HG** | 1 | 29 | 0 |
| **ME** | 0 | 1 | 14 |

process. The EFA solutions are superior to those yielded by the wrappers or Relief. Among the classifiers, NN seems the best alternative in general.

How significant are the results for the NN classifier? Treating test set accuracy as a binomial random variable, and considering a gaussian approximation thereof, the following CIs at the 95% level can be derived: 75% – 95% for the NR case, at least 88% for BSS-L(1) and BSS-L(5) and 80% – 97% for FSS-NN. Note that the CIs are very wide, given the small size of the involved test sets.

In order to ascertain which super-classes are the most difficult to predict, the two test set confusion matrices are shown in Table 3 (center and right).

**Table 4.** Test set performance on the $^{1}$H-MRS short-echo time. NR = no reduction, $k$ = number of neighbours, $cp$ = complexity parameter, $nt$ = number of trees. Best result has been highlighted in bold.

| Reduction | BSS | NN | | | NB | | C4.5 | | | RF | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| method | size | Acc | $F_1$ | k | Acc | $F_1$ | Acc | $F_1$ | cp | Acc | $F_1$ | nt |
| NR | 195 | 83.5% | 81.0% | 3 | 77.6% | 72.4% | 73.1% | 67.8% | 0.25 | 85.0% | 81.9% | 14 |
| BSS-S(1) | 6 | 89.5% | 87.2% | 10 | **92.5%** | **90.7%** | 88.1% | 86.8% | 0.25 | 86.5% | 85.2% | 11 |
| BSS-S(2) | 6 | 86.5% | 83.5% | 5 | 89.5% | 87.3% | 88.1% | 86.8% | 0.25 | 88.0% | 85.5% | 13 |
| Relief | 21 | 85.0% | 81.5% | 7 | 86.5% | 84.2% | 79.1% | 75.5% | 0.25 | 85.0% | 83.0% | 9 |

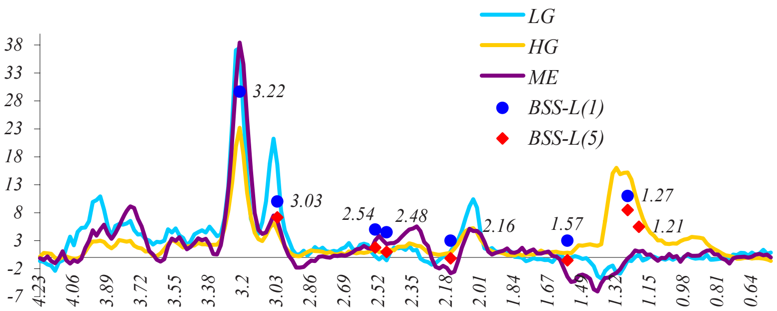


**Fig. 1.** The 7 spectral frequencies in EFA-L(1) and EFA-L(5)

**Table 5.** Left: $^1$H-MRS short-echo time test set results using Forward Subset Selection (FSS) in wrapper mode. Center and right: test set confusion matrices using NB on the subsets BSS-S(1) and BSS-S(2), respectively. True class falls vertically.

| method | size | Acc | $F_1$ | $k/cp$ |
|---|---|---|---|---|
| FSS-NN | 5 | 88.5% | 84.5% | 5 |
| FSS-NB | 11 | 88.0% | 86.6% | - |
| FSS-C4.5 | 12 | 83.6% | 79.8% | 0.25 |

| True | **BSS-S(1)+NB** | | |
|---|---|---|---|
| Class | **LG** | **HG** | **ME** |
| **LG** | 10 | 0 | 1 |
| **HG** | 3 | 35 | 0 |
| **ME** | 0 | 1 | 17 |

| True | **BSS-S(2)+NB** | | |
|---|---|---|---|
| Class | **LG** | **HG** | **ME** |
| **LG** | 10 | 0 | 1 |
| **HG** | 3 | 34 | 1 |
| **ME** | 1 | 1 | 16 |

Previous published work analyzing similar $^1$H-MRS data used PCA followed by LDA to distinguish between high-grade malignant tumours and meningiomas, obtaining a mean AUC (area under the ROC curve) of 0.94, using 6 principal components [4]. The same method was used to distinguish between high-grade malignant tumours and astrocytomas Grade II (part of the low-grade gliomas super-class), obtaining a mean AUC of 0.92, also using 6 principal components. There is an interesting link with the present work in the low number of final dimensions (six versus seven). However, two drawbacks of PCA are that all the spectra participate in the linear combination, and the fact that the linear combination may mix both positive and negative weights, which might partly cancel each other. In [2], LDA with 6 spectral frequencies (3.72, 3.04, 2.31, 2.14, 1.51 and 1.20 ppm) achieved a 83% of correct classification on an independent test set, this time using exactly the same three super-classes that we have analyzed in this study. Two possible reasons for the comparatively low performance may be that this particular solution does not make the problem linearly separable, or that the class distribution is markedly non-gaussian.

### 3.3 Short-Echo Time Experimental Results and Discussion

Short-echo time BSSs are reported in table 1 (right part). This time two subsets (named BSS-S(1) and BSS-S(2)) reached maximum relevance ($R = 1$), with just six spectral frequencies, graphically represented in Fig. 2 (note again the

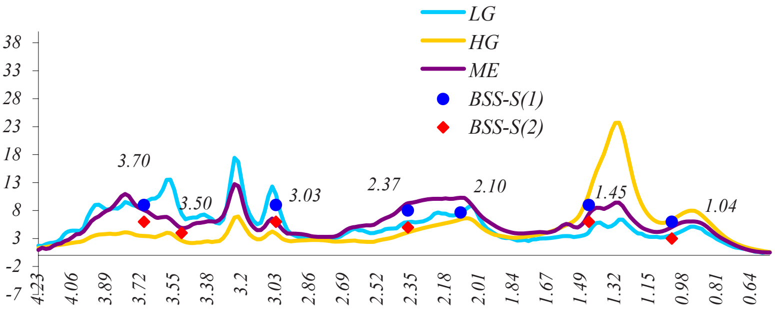


**Fig. 2.** The 6 spectral frequencies in EFA-S(1) and EFA-S(2)

Creatine peak at 3.03 ppm). The results for the four classifiers using the full set of frequencies, those obtained using BSS-S(1) and BSS-S(2) and Relief are displayed in Table 4, and those using Forward Selection in wrapper mode are shown in Table 5 (left). The two test set confusion matrices are shown in Table 5 (center and right). Again, the subsets BSS-S(1) and BSS-S(2) yielded by the EFA obtain the best results, above those with no reduction, Relief and the Forward Selection solutions. Contrary to the long-echo type of data, where clearly NN achieved the best results, this time the performance of NB is superior. Currently, we have no explanation for this discrepancy. The CIs at the 95% level are: 68% – 88% for the NR case, 86% – 99% for BSS-S(1) and BSS-S(2) and 81% – 96% for FSS-NN.

Previous existing work analyzing the same $^1$H-MRS data using 10 principal components and LDA as classifier obtained at most 85% of correct classification [3]. A different study reported 89% accuracy (again using LDA as classifier) with 5 spectral frequencies (3.76, 3.57, 3.02, 2.35, 1.28 ppm) in an independent test set [2]. These results are very much in agreement to those reported herein.

## 4 Conclusions and Future Work

Several feature selection methods have been applied to a high-dimensional $^1$H-MRS data set corresponding to different brain tumours. An entropic algorithm has been shown to be the best option, providing a drastic dimensionality reduction while improving on the performance of the full set of spectral frequencies. An added advantage of this method is its simplicity and the absence of any parameter tuning. Comparative results with similar studies show that the results are very competitive, both in terms of the prediction accuracy and the parsimony of the selected spectral frequencies, stressing the importance of feature selection algorithms in this particular kind of data.

Very noteworthy is the differential behaviour of the classifiers in presence of a strong reduction of dimension. In short-echo time data, the results for the NB and C4.5 classifiers are better in the reduced feature set, but worse in long-echo time data. The results for NN are always better and those of RF are about the same in both cases. The experimental results also seem to indicate that, contrary to previous work, long-echo time may yield better classification results, both before and after the feature selection process takes place, a fact that deserves further investigation. In particular, the introduction of resampling techniques as the bootstrap can yield better estimates. A specific and promising line of work will also study the combination of both echo types.

## Acknowledgements

Authors gratefully acknowledge the former INTERPRET partners (INTERPRET, EU-IST-1999-10310) and, from 1st January 2003, Generalitat de Catalunya (grants CIRIT SGR2001-194, XT 2002-48 and XT 2004-51); data providers: Dr. C. Majós (IDI), Dr.A. Moreno-Torres (CDP), Dr. F.A. Howe and Prof. J. Griffiths (SGUL), Prof. A. Heerschap (RU), Dr. W. Gajewicz (MUL) and Dr. J. Calvar (FLENI); data curators: Dr. A.P. Candiota, Ms. T. Delgado, Ms. J. Martín, Mr. I. Olier, Mr. A. Pérez and Prof. Carles Arús (all from GABRMN-UAB). Authors also acknowledge funding for CICyT TIN2006-08114 and SAF2005-03650 projects; the Mexican CONACyT and Baja California University and thank the anonymous reviewers for their helpful suggestions.

## References

1. Sibtain, N.: The clinical value of proton magnetic resonance spectroscopy in adult brain tumours. Clinical Radiology 62, 109–119 (2007)
2. Tate, A., et al.: Development of a decision support system for diagnosis and grading of brain tumours using in vivo magnetic resonance single voxel spectra. NMR in Biomedicine 19, 411–434 (2006)
3. Ladroue, C.: Pattern Recognition Techniques for the Study of Magnetic Resonance Spectra of Brain Tumours. PhD thesis, St. George's Hospital Medical School (2003)
4. Devos, A.: Quantification and classification of MRS data and applications to brain tumour recognition. PhD thesis, Katholieke Univ. Leuven (2005)
5. González, F., Belanche, L.: Feature selection in proton magnetic resonance spectroscopy for brain tumor classification. In: Procs. of ESANN 2008 (2008)
6. Wang, H.: Towards a unified framework of relevance. PhD thesis, Univ. of Ulster (1996)
7. Ng, M., Chan, L.: Informative gene discovery for cancer classification from microarray expression data. In: IEEE Workshop on Machine Learning for Signal Processing, pp. 393–398. IEEE, Los Alamitos (2005)
8. Le, D., Satoh, S.: Robust object detection using fast feature selection from hugh feature sets. In: 13th International Conference on Image Processing, pp. 961–964. IEEE, Los Alamitos (2006)
9. Kurgan, L., Cios, K.: Caim discretization algorithm. IEEE Transactions on Knowledge and Data Engineering 16(2), 145–153 (2004)
10. Majos, C., et al.: Brain tumor classification by proton mr spectroscopy: Comparison of diagnostic accuracy at short and long te. American Journal of Neuroradiology 25, 1696–1704 (2004)

11. Garcia, J., et al.: On the use of long te and short te sv mr. spectroscopy to improve the automatic brain. tumor diagnosis. Technical report (2007), ftp://ftp.esat.kuleuven.ac.be/pub/SISTA/ida/reports/07-55.pdf
12. Breiman, B.: Random forests. Machine Learning 45(1), 5–32 (2001)
13. Kira, K., Rendell., L.: The feature selection problem: Traditional methods and a new algorithm. In: Procs. of the Natl. Conf. on Artificial Intelligence, pp. 129–134 (1992)
14. Kohavi, R., John, G.: Wrappers for feature subset selection. Artificial Intelligence 97(1-2), 273–324 (1997)

# Statistical Assessment of MSigDB Gene Sets in Colon Cancer

Angela Distaso[1], Luca Abatangelo[1], Rosalia Maglietta[1],
Teresa Maria Creanza[1], Ada Piepoli[2], Massimo Carella[2],
Annarita D'Addabbo[1], Sayan Mukherjee[3], and Nicola Ancona[1]

[1] ISSIA-CNR, Via Amendola 122/D-I, 70126 Bari, Italy
[2] IRCCS-Casa Sollievo della Sofferenza Ospedale, San Giovanni Rotondo (FG), Italy
[3] Institute for Genome Sciences & Policy, Durham, NC 27708
{distaso,abatangelo,maglietta,creanza,daddabbo,ancona}@ba.issia.cnr.it

**Abstract.** Gene expression profiling offers a great opportunity for understanding the key role of genes in alterations which drive a normal cell to a cancer state. A deep understanding of the mechanisms of tumorigenesis can be reached focusing on deregulation of gene sets or pathways. We measure the amount of deregulation and assess the statistical significance of predefined pathways belonging to MSigDB collection in a colon cancer data set. To measure the relevance of the pathways we use two well-established methods: Gene Set Enrichment Analysis (GSEA) [7] and Gene List Analysis with Prediction Accuracy (GLAPA) [8]. We found that pathways associated to different diseases are strictly connected with colon cancer. Our study highlights the importance of using gene sets genes for understanding the main biological processes and pathways involved in colorectal cancer. Our analysis shows that many of the genes involved in these pathways are strongly associated to colorectal tumorigenesis.

**Keywords:** Microarray, pathway analysis, prediction accuracy, machine learning.

## 1 Introduction

Gene expression profiling has become a mainstay in the current research in applied genomics [1]. Current clinical practice consists in collecting specimens of tissues in two different phenotypical conditions, such as diseased patients vs. healthy controls. In this framework a major challenge is the identification of the main pathways or biological processes involved in the analyzed pathology. Such processes are coded through lists of genes defined on the basis of a-priori biological knowledge or experimentally. In the first case, such lists are composed of those genes which cooperate or are co-expressed in a particular cellular mechanism or function [3,4,5]. In the second case, the gene set represents the signature (response) of cells (system) to a given stimulus [6]. The focus of the problem is to identify biological processes, cellular functions and pathways perturbed in the phenotypic conditions by analyzing genes co-expressed in a given

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 206–213, 2008.

pathway as a whole, taking into account the possible interactions among them and, more important, the correlation of their expression with the phenotypical conditions [6,9].

In this paper we describe the results obtained by applying this approach to a data set composed of gene expression profiles relative to a case-control study of patients affected by colon cancer. Two well known methods recently proposed for finding deregulated pathways were applied. GSEA (Gene Set Enrichment Analysis) [7] finds perturbed pathways comparing the rank of genes in the data set with the ones belonging to the given pathway. To this end a Kolmogorov-Smirnov like statistic is used for assessing the statistical significance of the deregulation. GLAPA (Gene List Analysis with Prediction Accuracy) [8] uses the prediction accuracy of the phenotypic status of the patients for finding the pathways involved in the pathology. Both use non parametric permutation tests [12] and false discovery rate (FDR) [2] for assessing the statistical significance of the estimates. The database of gene sets we use in this study is the Molecular Signatures Database (MSigDB) [7]. This is a collection of 1687 curated gene sets with sizes ranging from 2 to 1594 genes, obtained from online pathway databases, publications in PubMed and expert knowledge.

## 2 Materials and Methods

### 2.1 Data Set Description

The gene expression profile data set was collected in Casa Sollievo della Sofferenza Hospital, Foggia - Italy [11]. The data set is made up of 22 normal and 25 tumor specimens of patients affected by colon cancer, profiled using the Affymetrix (Santa Clara, CA) HGU133A GeneChip (22283 probe-sets).

### 2.2 GLAPA: Gene List Analysis with Prediction Accuracy

GLAPA method has been proposed by any of the co-authors in [8]. The method uses a novel approach for finding deregulated gene sets. As a measure of relevance or deregulation of a given gene set, the prediction accuracy of the phenotype is used. The rationale is that a functional category coded through a list of genes is perturbed in a particular disease if it is possible to correctly predict the occurrence of the pathology in new subjects on the basis of the expression levels of those genes only. In other words, a functional category is informative for or is deregulated in a disease if the expression levels of the genes involved in the category are useful for training classifiers able to generalize, that is, able to correctly predict the status of new subjects [26]. So, generalization ability of predictors trained by using the expression levels of the genes co-operating in a given cellular mechanism or function can be seen as a measure of the relevance of the function in the pathology at hand. With the aim of estimating the relevance of a given pathway L we compute the prediction error $e_L$ of a linear Regularized Least Squares (RLS) classifier [27]. We trained the classifier by using the genes in the gene set with the strategy described in [8]. In particular we evaluated the error rate associated to the gene set performing the leave-k-out cross validation (LKOCV) strategy. The statistical significance of the measured accuracy is assessed against a couple of null hypothesis by using two independent permutation tests [12]. The first one (T1) aims at measuring how $e_L$ is due to the actual correlation between the genes in $L$ and the phenotype and how it is due by chance. To this end, we estimate the empirical probability density function of $e_L$ under the null hypothesis $H_0^y$ in which the genotype and the phenotype are supposed to be independent random variables. The nominal $P$-value $p_y$ relative to $e_L$ is so given by the percentage of random errors smaller than $e_L$. The second permutation test (T2) aims at evaluating how $e_L$ is dependent on the $n$ genes cooperating in the biological function coded by the list $L$ and how it depends only on the size of the list. In particular, in this test we assess if lists of the same size as $L$, composed of genes randomly selected from the ones present on the microarray, produce error rates smaller than $e_L$. To this end, we estimate the empirical probability density function of $e_L$ under a different null hypothesis $H_0^n$, where $n$ denotes the size of $L$. Under this hypothesis, we assume that *any* set $L^*$ of $n$ probes provides an error rate less than or equal to $e_L$ for predicting the *actual* phenotypic labels. The nominal $P$-value $p_n$ relative to $e_L$ is estimated as the percentage of random errors smaller than $e_L$. Moreover to account for multiple hypothesis testing, an estimate of the False Discovery Rate (FDR) [2] is computed. FDR is defined by the proportion of false hypothesis findings over the amount of alternative hypotheses accepted at a given level of statistical significance. The same procedure is adopted for computing an estimate of $\text{FDR}_y$ relative to $p_y$ and an estimate $\text{FDR}_n$ relative to $p_n$. Another parameter that we have considered in evaluating the tests is the power ($\pi_y$ and $\pi_n$), defined as the probability of accepting the alternative hypothesis $H_1$ when it is true.

### 2.3 GSEA: Gene Set Enrichment Analysis

GSEA [7] provides a statistical method for assessing the significance of predefined gene-sets starting from a microarray experiment. Given a gene expression dataset, the genes are ordered in a ranked list S according to their differential expression between the two classes. GSEA provides a score which measures the degree of enrichment of a given gene-set L at the extremes (top or bottom) of the rank-ordered list S. GSEA is based on a maximum deviation statistic of two distribution functions, similarly to the Kolmogorov-Smirnov test that is used to estimate the difference between the distributions. In fact, the score is calculated by walking down the list S, increasing a running-sum statistic when a gene in the gene-set in encountered, and decreasing it when genes not belonging to the gene-set are encountered. The magnitude of the increment depends on the correlation of the gene with the phenotype. The enrichment score (ES) is the maximum deviation from zero in the walk. The gene-sets related to the phenotypic distinction will tend to show high values of the ES. The significance of the ES is assessed by permutation testing: the observed ES is compared with the distribution of enrichment scores under the null hypothesis that the genotype and the phenotype are supposed to be independent random variables. The nominal p-value is given by the percentage of random enrichment scores greater than the observed value of ES. This procedure is similar to the one performed in the T1 permutation test of GLAPA.

Note that, with the aim of comparing analysis results across gene sets, the primary statistic suggested by the authors of GSEA, is the normalized enriched scores (NES). In fact by normalizing the ES, GSEA takes into account the differences in gene set size and correlations between gene sets and the expression data sets. NES is based on the gene sets enrichment scores for all dataset permutations. Hence in our experiment we refer to NES for examining the relevance of the gene sets.

## 3 Results and Discussion

### 3.1 Statistical Analysis

The deregulation of the whole collection of gene sets belonging to MSigDB was measured applying GSEA and GLAPA tools on our colon cancer data set independently. The GSEA software parameters were set to their default values. The statistical significance of normalized enrichment score (NES) associated to each gene set was assessed through a non parametric permutation test in which 1000 random permutations of the phenotypic labels were carried out. GSEA found 915 gene sets up-regulated in tumor and 769 up-regulated in normal specimens. Among these, only 399 gene sets up-regulated in tumor and 3 up-regulated in normal were found statistically significants with $\widehat{\mathrm{FDR}}_y \leq 25\%$.

For measuring the deregulations of each gene set $L$ with GLAPA, we measured the prediction error of the phenotype $e_L$ associated to $L$. To this end, for each gene set, 1000 cross validations of the data set were carried out. In each cross validation, we used 30 examples for training and the remaining 17 for testing RLS classifiers with linear kernel. We found 1381 pathways with an error rate $e_L \leq 25\%$. For assessing the statistical significance of $e_L$ with the permutation test T1, 1000 random permutations of the phenotypic labels were performed. This permutation test revealed 690 statistically significant gene sets ($p_y \leq 0.01$, $\widehat{\mathrm{FDR}}_y \leq 0.024$) having error rates $e_L \leq 17\%$. In order to determine if the deregulation of a particular pathway was due to the identity of the genes cooperating in the given pathway, or simply to the number of genes present in the gene set, the permutation test T2 was carried out. Specifically, 1000 gene sets were generated composed of $n$ probes randomly drawn from the ones available on the microarray. The error rate associated to each random gene set was evaluated performing 200 cross validations and compared with the error rate $e_L$. Such analysis revealed 58 pathways ($p_n \leq 0.02$, $\widehat{\mathrm{FDR}}_n \leq 0.25$) having an error rate $e_L \leq 11\%$ ($p_y \leq 0.010, \widehat{\mathrm{FDR}}_y \leq 0.024$). The table 1 shows the 21 statistically significant pathways found deregulated by both methods.

### 3.2 Biological and Functional Analysis

Here we focus on the $HDACI\_COLON\_SUL12HRS\_UP$ gene set that was found deregulated by GLAPA software only. Our aim is to find biological confirmation of its statistical significance. This gene set, composed of twenty six genes, seems to be specific for colorectal cancer. It was obtained experimentally by SW620 colonic epithelial cells as described in [13]. **ANXA2** (annexin A2; Location: 15q21-q22) and **ANXA5** (annexin A5; Location: 4q28-q32)genes encode two members of the annexin family, a calcium-dependent phospholipid-binding protein family that play a role in the regulation of cellular growth and in signal transduction pathways. It has been suggested that annexin 2 and annexin 5 are involved in cell proliferation/ differentiation and the pathogenesis of carcinoma. Their overexpression has been reported in various carcinomas including colon malignant tumors, and some findings suggest that they may be related to the progression and metastatic spread of colorectal carcinoma [14].

The second gene analyzed was **API5** (apoptosis inhibitor 5; Location: 11p11.2) gene that is an inhibitor of apoptosis. Many growth factors and cytokines act as cellular survival factors by preventing programmed cell death. Its expression is often upregulated in tumor cells, particularly in metastatic cells; so inhibition of Api5 function might offer a possible mechanism for antitumor exploitation [15].

Another gene belonging to this list is decay-accelerating factor (DAF) **CD55** (decay accelerating factor for complement; Location: 1q32) that is a membrane glycoprotein that regulates complement activation. The expression of DAF is enhanced in colorectal cancer cells and the colonic epithelium of ulcerative colitis in relation to the degree of mucosal inflammation [16] [17].

The **CDH1** (cadherin 1, type 1, E-cadherin (epithelial); Location: 16q22.1) gene belongs to the cadherin superfamily. The encoded protein is a calcium dependent cell-cell adhesion glycoprotein. Mutations in this gene are correlated with gastric, breast, colorectal, thyroid and ovarian cancer. The examination of E-cadherin expression and distribution in colorectal tumors can be extremely valuable in predicting disease recurrence [10].

Another important gene analyzed was **GSR** (glutathione reductase; Location: 8p21.1). The gastrointestinal tract is particularly susceptible to reactive oxygen species attack which lead to carcinogenesis. An important role in defense strategy against reactive oxygen species is played by antioxidants. In fact, superoxide dismutase, glutathione peroxidase and reductase are in general over-expressed in colorectal tumor [18].

**HSP90AA1** (heat shock protein 90kDa alpha (cytosolic), class A member 1; Location: 14q32.33) gene is a member of molecular chaperones, specifically of the heat shock protein 90 (Hsp90) family. HSP90 was low or non-detectable in normal colon tissues while high levels of HSP90 expression were observed in human colon cancer tissues, confirming the role of HSP90 as a potential marker for malignant colon cancer [19].

The protein encoded by **MYC** (v-myc myelocytomatosis viral oncogene homolog (avian); Location: 8q24.21) gene is a multifunctional, nuclear phosphoprotein that plays a role in cell cycle progression, apoptosis and cellular

**Table 1.** Pathways of MSigDB database deregulated in our colon cancer gene espression data set. For each pathway we report the name, the number of probes (size) and the most relevant statistical parameters as measured by GLAPA and GSEA tools.

| | | GLAPA | | | | | | | GSEA | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Pathway | Size | $e_L$ | $p_y$ | $FDR_y$ | $\pi_y$ | $p_n$ | $FDR_n$ | $\pi_n$ | NES | $p_y$ | $FDR_y$ |
| ADIPOCYTE BRCA UP | 17 | 0.09 | 0.003 | 0.02 | 0.97 | 0.008 | 0.12 | 0.68 | 1.7 | 0.016 | 0.04 |
| ADIP DIFF CLUSTER3 | 68 | 0.08 | 0.001 | 0.02 | 0.99 | 0.002 | 0.04 | 0.78 | 1.5 | 0.032 | 0.15 |
| AS3 HEK293 DN | 20 | 0.10 | 0.003 | 0.02 | 0.96 | 0.008 | 0.12 | 0.72 | 1.7 | 0.006 | 0.05 |
| BLEO HUMAN LYMPH HIGH 4HRS UP | 33 | 0.09 | 0.001 | 0.02 | 0.97 | 0.007 | 0.12 | 0.73 | 1.4 | 0.129 | 0.22 |
| BLEO MOUSE LYMPH HIGH 24HRS DN | 79 | 0.11 | 0.002 | 0.02 | 0.97 | 0.015 | 0.21 | 0.61 | 1.6 | 0.052 | 0.08 |
| CANCER UNDIFFERENTIATED META UP | 96 | 0.09 | 0.002 | 0.02 | 0.98 | 0.008 | 0.12 | 0.69 | 1.9 | 0.001 | 0.01 |
| CARBON FIXATION | 37 | 0.09 | 0.009 | 0.02 | 0.84 | 0.008 | 0.12 | 0.74 | 1.59 | 0.029 | 0.10 |
| CELL CYCLE CHECKPOINT | 47 | 0.11 | 0.004 | 0.02 | 0.97 | 0.02 | 0.24 | 0.60 | 1.81 | 0.008 | 0.03 |
| CIS RESIST LUNG UP | 19 | 0.11 | 0.004 | 0.02 | 0.95 | 0.02 | 0.24 | 0.66 | 1.55 | 0.049 | 0.12 |
| CMV HCMV TIMECOURSE 14HRS UP | 77 | 0.11 | 0.001 | 0.02 | 0.96 | 0.02 | 0.24 | 0.64 | 1.61 | 0.021 | 0.09 |
| CROONQUIST IL6 RAS DN | 37 | 0.10 | 0.001 | 0.02 | 0.97 | 0.02 | 0.24 | 0.66 | 1.71 | 0.018 | 0.05 |
| FATTY ACID SYNTHESIS | 26 | 0.11 | 0.003 | 0.02 | 0.97 | 0.02 | 0.24 | 0.62 | 1.54 | 0.048 | 0.13 |
| HG PROGERIA DN | 39 | 0.11 | 0.002 | 0.02 | 0.97 | 0.02 | 0.24 | 0.67 | 1.84 | 0.004 | 0.02 |
| HIFPATHWAY | 31 | 0.10 | 0.002 | 0.02 | 0.97 | 0.01 | 0.13 | 0.69 | 1.69 | 0.013 | 0.06 |
| JAIN NEMO DIFF | 125 | 0.09 | 0.001 | 0.02 | 0.99 | 0.01 | 0.05 | 0.72 | 1.86 | 0.002 | 0.02 |
| KLEIN PEL UP | 102 | 0.09 | 0.002 | 0.02 | 0.99 | 0.01 | 0.09 | 0.71 | 1.73 | 0.006 | 0.05 |
| NEMETH TNF DN | 53 | 0.09 | 0.002 | 0.02 | 0.98 | 0.01 | 0.04 | 0.74 | 1.77 | 0.002 | 0.04 |
| PENTOSE PHOSPHATE PATHWAY | 35 | 0.07 | 0.002 | 0.02 | 0.93 | 0.001 | 0.001 | 0.85 | 1.61 | 0.025 | 0.09 |
| PURINE METABOLISM | 191 | 0.10 | 0.002 | 0.02 | 0.98 | 0.015 | 0.212 | 0.63 | 1.85 | 0.002 | 0.021 |
| SA G2 AND M PHASES | 16 | 0.11 | 0.003 | 0.02 | 0.96 | 0.016 | 0.225 | 0.65 | 1.70 | 0.025 | 0.056 |
| ZHAN MMPC SIM BC AND MM | 88 | 0.10 | 0.002 | 0.02 | 0.98 | 0.021 | 0.243 | 0.65 | 1.54 | 0.035 | 0.126 |

transformation. It functions as a transcription factor that regulates transcription of specific target genes. Mutations, overexpression, rearrangement and translocation of this gene have been associated with a variety of tumors. c-MYC protects from p53-mediated apoptosis. Some findings indicate that failure of the normal apoptotic process together with de-regulation of c-MYC proto-oncogene might promote the development of colorectal tumors and its overexpression is observed in most colorectal cancers [20] [21].

An essential requirement for the development, progression and metastasis of malignant tumors is angiogenesis. **VEGF** (vascular endothelial growth factor; Location: 6p12) plays an essential role in the development of angiogenesis of numerous solid malignancies, including colon cancer. This gene is a member of the PDGF/VEGF growth factor family and acts on endothelial cells mediating increased vascular permeability, inducing angiogenesis, vasculogenesis and endothelial cell growth, promoting cell migration, and inhibiting apoptosis. VEGF is associated with the development and prognosis of colorectal cancer, but its relation with degree of differentiation remains to be studied [22]; likely VEGF functions as regulators of colon cancer cell invasion. VEGF expression is induced in colon and other cancer cells as a result of hypoxia and multiple genetic alterations. However it is evident that there is an association between VEGF expression, p53 status and angiogenesis, suggesting that mutant p53 plays a central role in promoting angiogenesis in colon cancer progression [23].

## 4 Conclusions

In this paper we have described the biological and functional relevance in colon cancer of pathways found deregulated in a gene expression profile data set relative to a case-control study of patients affected by this pathology [11]. GLAPA and GSEA methods were applied for measuring deregulation of pathways and for assessing their statistical significance [7,8]. The pathway database used in this study is a curated collection of 1687 gene sets obtained from different sources [7]. Other studies have pointed out the fundamental role of pathways in studying onset and progression of tumors [6,24]. Indeed cancer is a heterogeneous disease caused by a complex of altered processes that could be grouped according to the six hallmarks of cancer (self-sufficiency in growth signals, insensitivity to anti-growth signals, evasion of apoptosis, limitless replicative potential, sustained angiogenesis, tissue invasion and metastasis) [25]. Our study highlights that pathway approach to the investigation of complex diseases allows to get a well comprehensive picture of altered biological processes in cancer pathology.

## References

1. Schena, M., et al.: Science (270), 467–470 (1995)
2. Storey, J.D., et al.: Proc. Natl. Acad. Sci. (100), 9440–9445 (2003)
3. Ashburner, M., et al.: Nat. Genet. (25), 25–29 (2000)
4. Kanehisa, M., et al.: Nucleic Acids Res. (30), 42–46 (2002)
5. Khatri, P., et al.: Genomics 79(2), 266–270 (2002)
6. Bild, A.H., et al.: Nature 19(439), 353–357 (2006)
7. Subramanian, A., et al.: Proc. Natl. Acad. Sci. (102), 15545–15550 (2005)
8. Maglietta, R., et al.: Bioinformatics 23(16), 2063–2072 (2007)
9. Creighton, C.J., et al.: PLoS ONE 3(3), 1816 (2008)
10. Elzagheid, A., et al.: World J. Gastroenterol. 12(27), 4304–4309 (2006)
11. Ancona, N., et al.: BMC Bioinformatics 387(7) (2006)
12. Good, P., et al.: Springer, New York (1994)
13. Mariadason, J.M., et al.: Cancer Res. 60(16), 4561–4572 (2000)
14. Emoto, K., et al.: Cancer (92), 1419–1426 (2001)
15. Morris, E.J., et al.: PLoS Genetics 2(11), 1834–1848 (2006)
16. Okazazi, H., et al.: J. Lab. Clin. Med. 143(3), 169–174 (2004)

17. Durrant, L.G., et al.: Cancer Immunol. Immunother. (52) (2003)
18. Skrzydlewska, E., et al.: Hepatogastroenterology 50(49), 126–131 (2003)
19. Park, K.A., et al.: Carcinogenesis 28(1), 71–80 (2007)
20. Greco, C., et al.: Anticancer Res. 21(5), 3185–3192 (2001)
21. Seidler, H.B.K., et al.: Experimental and Molecular Pathology (76), 224–233 (2004)
22. Han, J., et al.: Zhonghua Yi Xue Za Zhi 82(7), 481–483 (2002)
23. Faviana, P., et al.: Oncol. Rep. 9(3), 617–620 (2002)
24. Edelman, E., et al.: PLoS Computational Biology 4(2), 28 (2008)
25. Hanahan, D., et al.: Cell (100), 57–70 (2000)
26. Vapnik, V.: The Nature of Statistical Learning Theory. Springer, New York (1995)
27. Rifkin, R., Yeo, G., Poggio, T.: Advances in Learning Theary: Methods, Model and Applications. In: Suykens, H., Basu, M., Vandewalle, A. (eds.) NATO Science Series III: Computer and Systems Science, vol. 190, pp. 131–153. IOS Press, Amsterdam (2003)

# Stratification of Severity of Illness Indices: A Case Study for Breast Cancer Prognosis

Terence A. Etchells[1], Ana S. Fernandes[2], Ian H. Jarman[1],
José M. Fonseca[2], and Paulo J.G. Lisboa[1]

[1] School of Computing and Mathematical Sciences, Liverpool John Moores University,
Byrom Street, Liverpool L3 3AF, UK
{T.A.Etchells, I.H.Jarman, P.J.Lisboa}@ljmu.ac.uk
[2] Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa
{asff, jmf}@uninova.pt

**Abstract.** Prognostic modelling involves grouping patients by risk of adverse outcome, typically by stratifying a severity of illness index obtained from a classifier or survival model. The assignment of thresholds on the risk index depends of pairwise statistical significance tests, notably the log-rank test. This paper proposes a new methodology to substantially improve the robustness of the stratification algorithm, by reference to a statistical and neural network prognostic study of longitudinal data from patients with operable breast cancer.

**Keywords:** Risk modelling, log-rank test.

## 1 Introduction

Stratification of patients by risk of adverse outcome is central to clinical practice. This begins with modelling empirical data either with a classifier or a failure time model, depending on whether the data represent a snapshot in time of the patient's condition at diagnosis, or evolution of the disease over time in a longitudinal cohort study. Either way, the equivalent of the linear argument *β.x* in a Generalised Linear Model defines a prognostic index that ranks patient data by severity of the illness. In the case of breast cancer, typically a piecewise linear model is used [1] from which the prognostic index can be derived. A good example of this is the Nottingham Prognostic Index (NPI) which is widely used in clinical practice [2] and takes the form: *NPI score = 0.2*Tumour size (cm) + Node Stage (1…3) + Histological Grade (1…3).*

The same principles apply when flexible models are used, such as generic non-linear algorithms including artificial neural networks.

In the case of discrete time models of longitudinal data, the main variable is the event rate, also called the hazard rate

$$h(x_p, t_k) = P(t \le t_k \mid t > t_{k-1}, x_i) \tag{1}$$

which is the probability that patient $p$ with characteristics $x_p$ survives to the end of time interval $t_k$ given that the patient is known to have entered that interval without

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 214–221, 2008.

experiencing the event of interest. This is the output of the Partial Logistic Artificial Neural Network (PLANN) model [3]. This index is usually averaged over time to define a time-independent risk score over a predefined time interval.

The equivalent index to the NPI score is now the log-odds ratio

$$PLANN\ Risk\ Index = \log\left(\frac{h(x_p,t_k)}{\left(1\text{-}h(x_p,t_k)\right)}\right) \tag{2}$$

where the conditional probability of class membership is directly estimated by the neural network output.

Once the risk score is defined, the population of patients at risk needs to be stratified for the purpose of tailoring adjuvant therapy and to enable comparisons between to be made between patient cohorts from different clinical centres, or subject to different clinical interventions, to be made between patients at similar risk by outcome. In survival analysis the most widely used statistic to test significant differences is the log-rank statistic.

The next section reviews current practice in the application of the log-rank test to stratification of patient data. This is followed by a case study of prognostic modelling of data from patients with operable breast cancer, comparing a statistical methodology (Cox regression) with PLANN-ARD [4] that uses the ARD framework to regularise the PLANN model and performed well in comparison with alternatives in a recent double blind benchmark of linear and generic non-linear survival models [5]. In section 4 the prognostic scores obtained on the same data by these two methods are stratified, comparing the standard application of the log-rank test with a novel methodology proposed in this paper, to resolve significant issues of robustness in the allocation of patients into risk groups.

## 2 Application of the Log-Rank Test to Stratification of Patient Data

In the literature the approach to splitting risk indices into risk groups is not always stated clearly, sometimes stating the cut-off points of the respective risk scores without a clear indication of how these were obtained [6-7]. Where the split of the indices is at all explained, expert knowledge has been a factor as in the case for the widely used NPI. This index is designed for ease of use and is derived by rounding a more cumbersome Cox regression calculation, the cut-off points being chosen to best match the risk groups from the original model which, in turn, was split on the basis of best match with known clinical groups.

In another approach the indices are split into equal sized groups as suggested by Harrell et al [8]. This tutorial in biostatistics suggests using deciles as a starting choice and in a prognostic model for ovarian cancer Clark et al [9] used quartiles to partition the risk score.

A suggestion for an automated method is to use successive top-down splits by maximising the log-rank test statistic [10].

## 3 Prognostic Modelling of Breast Cancer Patients

The reference data set for this case study consists of routinely acquired clinical records for patients recruited by Christie Hospital, Manchester, during 1983-89. The specific cohort of interest is patients with early, or operable breast cancer, defined using the standard TNM (Tumour, Nodes Metastasis) staging system as tumour size less than 5 cm, node stage less than 2 and without clinical symptoms of metastatic spread. This defines a case series (n=917) for a longitudinal cohort study with 5-year follow-up. The date of recruitment is the date of surgery and the event of interest is cancer specific mortality.

Earlier studies identified the following six predictive variables: age at diagnosis, node stage, histological type (lobular, ductal or in situ), ratio of axillary nodes affected to axilar nodes removed, pathological size (i.e. tumour size in cm) and oestrogen receptor count. All of these variables are banded and binary coded as 1-from-N. For details of the attribute assignment see [4].

Two analytical models were fitted to the data, starting with a piecewise linear model Cox regression, also known as proportional hazards. This model factorises dependence on time and the covariates, modelling the hazard rate for patient with clinical characteristics $x_p$ at time $t_k$ as follows:

$$\frac{h(x_p,t_k)}{1-h(x_p,t_k)} = \frac{h_0(t_k)}{1-h_0(t_k)}.\exp\left(\sum_{i=1}^{N_i}\beta x_i\right) \tag{3}$$

where $h_0$ denotes the empirical hazard for a reference population with covariate attributes all equal to zero and $x_i$ is the static covariate vector. This was taken to be the standard used by the software that implemented the piecewise linear model Cox regression (SAS), which is the last attribute for each covariate. It was found that the risk group allocation is not sensitive to the choice of reference population. In contrast, the PLANN model is semi-parametric and with the following model for the hazard.

$$\frac{h(x_p,t_k)}{1-h(x_p,t_k)} = \exp\left(\sum_{h=1}^{N_h} w_h \cdot \; g(\sum_{i=1}^{N_i} w_{ih}.x_{pi} + w.t_k + b_h) \;\; + b\right) \tag{4}$$

where the indices *i* and *h* denote the input and hidden node layers and the non linear function $g(\cdot)$ is a sigmoid.

Both models optimise the same objective function, namely the log-likelihood summed over the observed status of the patient sampled over of 60 months, with an indicator variable that is 1 for death attributed to breast cancer and 0 if the patient is observed alive. Using target values $\tau_{pk}$ as indicator labels and $t_l$ as the time index

$$G = -\sum_{p=1}^{No.patients}\sum_{k=1}^{t_l}\left[\tau_{pk}\log\left(h\left(x_p,t_k\right)\right)+\left(1-\tau_{pk}\right)\log\left(1-h\left(x_p,t_k\right)\right)\right] \tag{5}$$

Data for patients who are lost to follow-up are said to be right-censored and the patients no longer counts as part of the set of patients at risk.

In addition, over-fitting is avoided by regularisation, in the case of Cox regression using Akaike's Information Criterion (AIC) and in the neural network model with Automatic Relevance Determination (ARD) [11].

The prognostic index is defined in both cases as the covariate dependent term in eq. (2) and they are compared for the two models in fig. 1.

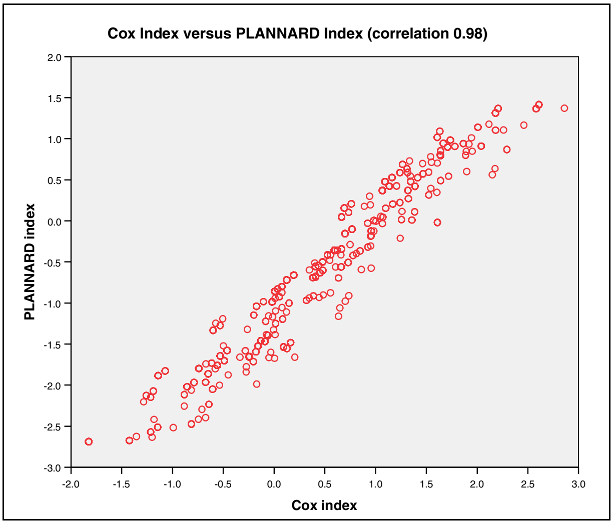


**Fig. 1.** Correlation between the prognostic derived with Cox regression and PLANN-ARD for a 5-year study of a patient cohort with early breast cancer. The proportionality of the hazards is borne out by the high correlation between the methods. In general, higher risk cohorts, longer follow-up times, studies of recurrence and models for other diseases will only be suitable for piecewise linear modelling if the proportionality of the hazards is observed.

The next stage in the modelling process is to use the prognostic index as the basis to partition the patient cohort into groups at similar risk of adverse outcome. This is the subject of the next section.

## 4 Robust Methodology for Stratification of Severity of Risk

The most widely used method for stratification of an empirical distribution of prognostic indices is to apply the log-rank test statistic from which the statistical significance for pairwise data partitions can be measured. Given that the test only applies in a pairwise manner, that is to say, for separating two cohorts at a time, this requires a search for the most appropriate threshold to divide the distribution of prognostic index scores.

An accepted strategy was implemented in SAS. It starts by sorting all the records by the value of the prognostic index. Next, the total data are divided into two groups at a threshold value that sweeps the full range of prognostic indices from minimum to

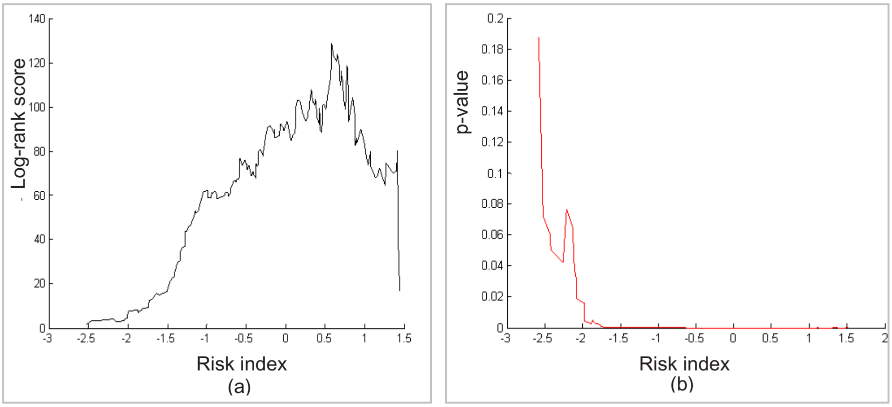


**Fig. 2.** The significance of data partitions in the top-down approach that is generally applied to stratify patient data in medical statistics detects the global maximum in (a), but this does not take into account that the statistical significance is high for a wider range of possible cut-off thresholds as shown in (b)

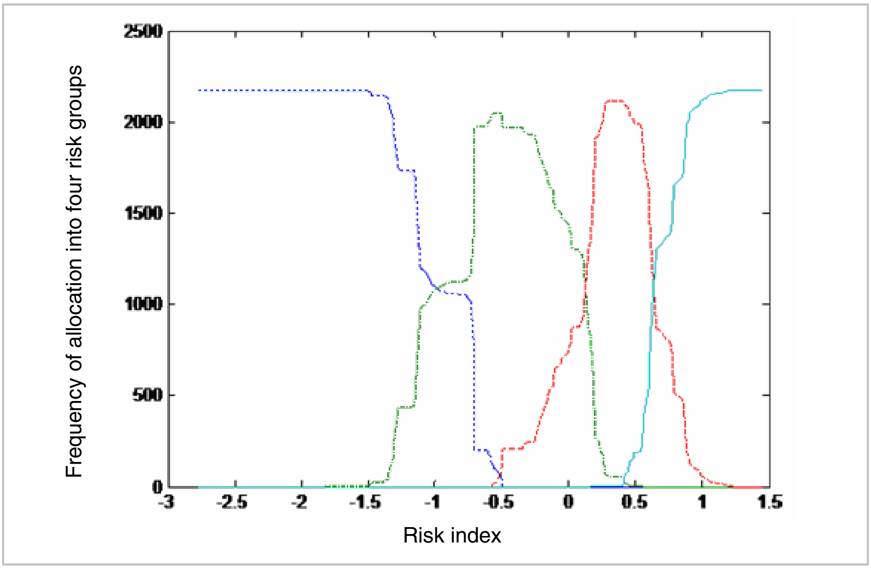


**Fig. 3.** Results from the proposed methodology for allocating risk groups from a risk index for severity of illness. Note that the group allocation frequencies vary smoothly, in contrast with the spot values of the log-rank test statistic in fig. 2.

maximum. For each threshold, the *log-rank statistic* is calculated and hence a *p-value* results. The maximum of the log-rank statistic determines the first cut-off point. The same method is then repeated in each of the separated cohorts until no further

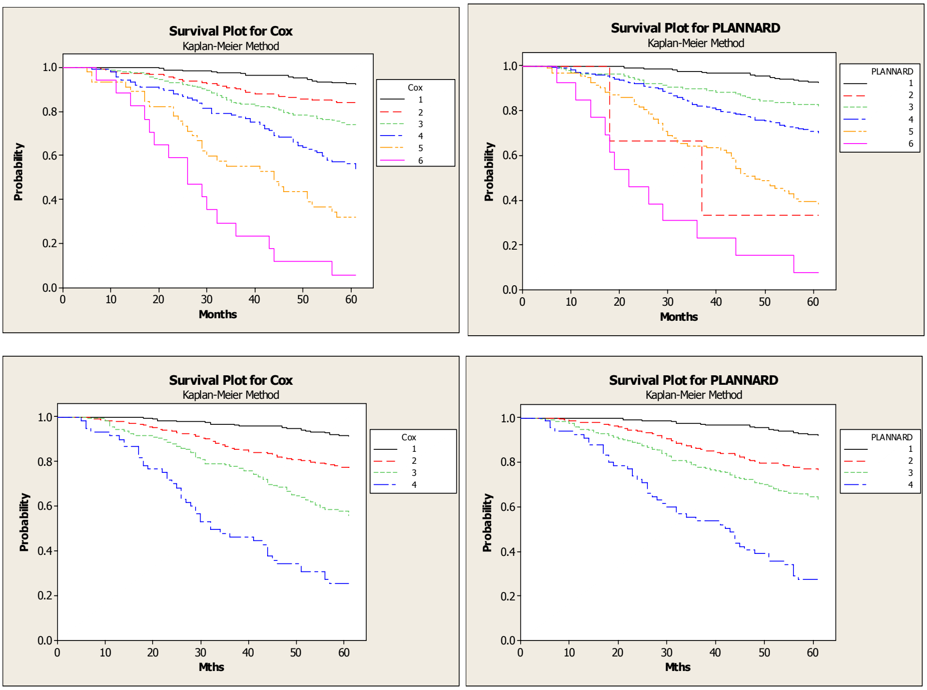


**Fig. 4.** Actuarial estimates of survival obtained with the Kaplan-Meier method, for the same cases (n=917), stratified using the log-rank test over a 60 month period. The top row uses the standard method and the bottom row uses the proposed method for increasing robustness in the risk stratification. The left column uses Cox regression modelling and the right column the PLANN-ARD neural network. The two modelling algorithms should be consistent, shown in fig. 1 but this is only apparent when the bootstrap method was applied.

partitioning exceeds a pre-set confidence level which, for this study, is as p-value of 0.01 (99% of confidence), corresponding to a test statistic value of around seven.

In practice, the test statistic very much exceeds this value across a wide range of thresholds with the associated p-values forming a plateau indicating that there are a wide range of candidate cutpoints in addition to the maximum log rank statistic that has been selected as can be seen in fig. 2.

A new methodology is proposed to make the stratification of risk indices more robust. The new approach is bottom-up according to the following procedure which involves two nested loops:

*Inner loop*

i. Bin the risk indices into discrete intervals each containing a minimum number of cases (e.g. $n_{min}$=10).
ii. Calculate the log-rank statistic for each pair of adjacent cells and aggregate together the two cells with the smallest value of this test statistic.

iii. Repeat the process until the long-rank statistic is significant for all remaining cell pairs. *Outer loop*

i. Draw a sample of the risk indices, with replacement, of size equal to the original data size – this is a bootstrap re-sample of the data.
ii. Apply the *inner loop* to convergence using the re-sampled data.
iii. Allocate each value in the full range of the risk index to a risk group, from $1..N_{groups}$
iv. Repeat from *i.* a given number of times (e.g. $n_{resamples}$ =3000)
v. Identify the distribution of values of $N_{groups}$ and discard all group assignments different from the mode of this distribution.
vi. For each value in the full range of the risk index, build a distribution of risk group allocations – this clearly indicates the cases that fit firmly into a risk group and those that are near the boundary between adjacent groups.
vii. Allocate each case in the original sample to the mode of the distribution of risk groups.

In the current case study, the risk group distributions obtained by this method are plotted in fig. 3.

The robustness of this approach to risk group identification is illustrated in fig. 4. The small size sample causing the unexpected outcome profiles in the solution with 6 risk groups may be an indication that this methodology is over-fitted to the training data.

## 5 Conclusions

The application of the log-rank test statistic to stratify patients by risk of adverse outcome is subject to variability due to the high prevalence of similar scores for many different risk thresholds. This results in unstable boundaries between strata, causing unwanted variability in allocation of patients into risk groups.

This paper proposes a robust methodology for risk group allocation which exploits bootstrap re-sampling in order to stabilise the distribution of risk groups predicted for each value of the risk score index. The effectiveness and robustness of this methodology are shown by reference to a case study for operable breast cancer, using data from a longitudinal cohort study with 5-year follow-up.

In addition, the generic applicability of the proposed methodology is illustrated using both piecewise linear and neural network models of survival. While the results are consistent with earlier studies of the same data, the current findings are regarded as definitive on account of the robustness that has been added to the stratification process.

**Acknowledgments.** This work was carried out with partial funding from the Biopattern Network of Excellence FP6/2002/IST/1; N° IST-2002-508803 (www.biopattern.org) and from the Fundação para a Ciência e Tecnologia, through the POS_C(SFRH/BD/30260/2006). Availability of data by Christie Hospital is gratefully acknowledged.

## References

1. Cox, D.R.: Regression models and life tables. Journal of the Royal Statistical Society, B 74, 187–220 (1972)
2. Haybittle, J.L., Blamey, R.W., Elston, C.W., Johnson, J., Doyle, P.J., Campbell, F.C., Nicholson, R.I., Griffiths, K.: A prognostic index in primary breast cancer. British Journal of Cancer 45, 3621 (1982)
3. Biganzoli, E., Boracchi, P., Mariani, L., Marubini, E.: Feed forward neural networks for the analysis of censored survival data: A partial logistic regression approach. Statistics in Medicine 17, 1169–1186 (1998)
4. Lisboa, P.J.G., Wong, H., Harris, P., Swindell, R.: A Bayesian neural network approach for modelling censored data with an application to prognosis after surgery for breast cancer. Artificial Intelligence in Medicine 28(1), 1–25 (2003)
5. Taktak, A., Antolini, L., Aung, M., Boracchi, P., Campbell, I., Damato, B., Ifeachor, E., Lama, N., Lisboa, P., Setzkorn, C., Stalbovskaya, V., Biganzoli, E.: Double-blind evaluation and benchmarking of survival models in a multi-centre study. Comput. Biol. Med. 37, 8 (2007)
6. Guerra, I., Algorta, J., Diaz de Otazu, R., Pelayo, A., Farina, J.: Immunohistochemical prognostic index for breast cancer in young women. J. Clin. Pathol: Mol. Pathol. 56, 323–327 (2003)
7. Ortiz Sebastian, S., Rodrıguez Gonzalez, J.M., Parilla Paricio, P., Sola Perez, J., Perez Flores, D., Pinero Madrona, A., Ramirez Romero, P., Tebar, F.J.: Papillary Thyroid Carcinoma: Prognostic Index for Survival Including the Histological Variety. Arch. Surg. 135 (March 2000)
8. Harrell, F.E., Lee, K.L., Mark, B.D.: Tutorial in Biostatistics Multivariate Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors. Statistics in Medicine 15, 361–387 (1996)
9. Clark, T.G., Stewart, M.E., Altman, D.G., Gabra, H., Smyth, J.F.: A prognostic model for ovarian cancer. Br. J. Cancer. 85, 944–952 (2001)
10. Williams, B.A., Mandrekar, J.N., Mandrekar, S.J., Cha, S.S., Furth, A.F.: Finding Optimal Cutpoints for Continuous Covariates with Binary and Time-to-Event Outcomes. Technical Report Series #79, Mayo Clinic, Rochester, Minnesota (June 2006)
11. MacKay, D.J.C.: Probable networks and plausible predictions – a review of practical Bayesian methods for supervised neural networks. Network: Computation in Neural Systems 6, 469–505 (1995)

# Categorization of Web Users by Fuzzy Clustering

Giovanna Castellano and Maria Alessandra Torsello

Computer Science Department, University of Bari,
Via E. Orabona, 4 - 70126 Bari, Italy
{castellano, torsello}@di.uniba.it
http://www.di.uniba.it/~cilab

**Abstract.** Categorization of users is a fundamental task in Web personalization. Fuzzy clustering is a valid approach to derive user categories by capturing similar user interests from web usage data available in log files. Usually, fuzzy clustering is based on the use of Euclidean metrics to evaluate similarity between user preferences. This can lead to user categories that do not capture the semantic information incorporated in the original Web usage data. To better capture similarity between users, in this paper we propose the use of a measure that is based on the evaluation of similarity between fuzzy sets. The proposed fuzzy measure is employed in a relational fuzzy clustering algorithm to discover clusters embedded in the Web usage data and derive categories modeling the preferences of similar users. An application example on usage data extracted from log files of a real Web site is reported and a comparison with the results obtained using the cosine measure is shown to demonstrate the effectiveness of the fuzzy similarity measure.

**Keywords:** Web mining, Fuzzy clustering, access log, Web personalization, user profiling.

## 1 Introduction

The rapid development of the Web as a new medium for information dissemination has given rise to an increasing interest for Web personalization. Web log files contain a huge amount of data about user access patterns. Hence, if properly exploited, they can reveal useful information about the browsing behavior of users in a site. As a consequence, these data can be employed to derive categories of users useful to deliver recommendations to currently connected users according to the discovered user categories.

To reveal information about user interests from Web log files, different works have proposed the application of Data Mining techniques, leading to the so-called Web Usage Mining (WUM). In general, a WUM approach applies Data Mining algorithms on Web usage data in order to discover interesting patterns in the user browsing behavior [7], [4]. Among data mining approaches, clustering is an effective way to group users with common browsing behavior [10], [11].

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 222–229, 2008.

In the choice of the clustering method for WUM, one important constraint to be considered is the possibility to obtain overlapping clusters, so that a user can belong to more than one group. Another key feature to be addressed is vagueness and imprecision inherent Web usage data. To deal with the ambiguity and the uncertainty underlying Web interaction data, as well as to derive overlapping clustering, fuzzy clustering appears to be an effective tool [8], [6], [9].

In this paper, we employ fuzzy clustering for the categorization of users visiting a Web site. In particular, we use CARD+, a fuzzy relational clustering algorithm that works on data quantifying similarity between user interests. Two main activities can be distinguished in our approach:

- The creation of the relation matrix containing the dissimilarity values among all pairs of users;
- The categorization of users by grouping similar users into categories.

Instead of using standard similarity measures, such as the cosine based similarity, we propose the use of a fuzzy measure to express similarity between Web users which is derived from the similarity quantification of fuzzy sets.

## 2 Creation of the Relational Data

The first activity in the categorization of similar users consists in the creation of the relation matrix including the dissimilarity values between all pairs of users. To create the relation matrix, it is essential to measure the similarity between two generic users on the basis of available data. Data about the browsing behavior of users are obtained by the preprocessing of log files storing all the information concerning the accesses to the Web site. The preprocessing of log files is performed by means of LODAP, a software tool that we have implemented for the analysis of Web log files in order to derive models characterizing the user browsing behaviors. This aim is achieved through four main steps: data cleaning, data structuration, data filtering and interest degree computation. Main details can be found in [3]. As a result, LODAP extracts data which are synthetized in a behavior matrix $\mathbf{B} = [b_{ij}]$ where the rows $i = 1, \ldots, n$ represent the users and the columns $j = 1, \ldots, m$ correspond to the Web pages of the site. Each component $b_{ij}$ of the matrix indicates the interest degree of the $i$-th user for the $j$-th page. The $i$-th user behavior vector $\mathbf{b}_i$ ($i$-th row of the behavior matrix) characterizes the browsing behavior of the $i$-th user. Based on the behavior matrix, the similarity between two generic users is expressed by the similarity between the two corresponding user behavior vectors.

In literature, different measures have been proposed to evaluate the similarity degree between two generic objects. One of the most common measures employed to this aim is the angle cosine measure. In the specific context of user category extraction, the cosine measure computes the similarity between any two behavior vectors $\mathbf{b}_x$ and $\mathbf{b}_y$ as follows:

$$Sim_{Cos}\left(\mathbf{b}_x, \mathbf{b}_y\right) = \frac{\mathbf{b}_x \mathbf{b}_y^{'}}{\|\mathbf{b}_x\| \, \|\mathbf{b}_y\|} = \frac{\sum_{j=1}^{m} b_{xj} b_{yj}}{\sqrt{\sum_{j=1}^{m} b_{xj}^2} \sqrt{\sum_{j=1}^{m} b_{yj}^2}}. \qquad (1)$$

The use of the cosine measure might be ineffective to define the similarity between two users visiting a Web site. In effect, to evaluate the similarity between two generic users (rows of the available matrix), the cosine measure takes into account only the common pages visited by the considered users. This approach may produce ineffective results, leading to the loss of semantic information underlying Web usage data related to the relevance of each page for each user.

To better capture the similarity between two generic Web users, we propose the use of a fuzzy similarity measure. Specifically, two generic users are modeled as two fuzzy sets and the similarity between these users is expressed as the similarity between the corresponding fuzzy sets. To do so, the user behavior matrix **B** is converted into a matrix $\mathbf{M} = [\mu_{ij}]$ which expresses the interest degree of each user for each page in a fuzzy way. A very simple characterization of the matrix **M** is provided as follows:

$$\mu_{ij} = \begin{cases} 0 & \text{if} \quad b_{ij} < ID_{min} \\ \frac{b_{ij} - ID_{min}}{id_{max} - ID_{min}} & \text{if}\; b_{ij} \in [ID_{min}, ID_{max}] \\ 1 & \text{if} \quad b_{ij} > ID_{max} \end{cases} \tag{2}$$

where $ID_{min}$ is a minimum threshold for the interest degree under which the interest for a page is considered null, and $ID_{max}$ is a maximum threshold of the interest degree, after which the page is considered surely preferred by the user.

Starting from this fuzzy characterization, the rows of the new matrix **M** are interpreted as fuzzy sets defined on the set of Web pages. Each fuzzy set $\mu_i$ is related to a user $\mathbf{b}_i$ and it is simply characterized by the following membership function:

$$\mu_i(j) = \mu_{ij} \qquad \forall j = 1, 2, \ldots, m \tag{3}$$

In this way, the similarity of two generic users is intuitively defined as the similarity between the corresponding fuzzy sets. The similarity among fuzzy sets can be evaluated in different ways [12]. One of the most common measures to evaluate similarity between two fuzzy sets is the following:

$$\sigma(\mu_1, \mu_2) = \frac{|\mu_1 \cap \mu_2|}{|\mu_1 \cup \mu_2|} \tag{4}$$

According to this measure, the similarity between two fuzzy sets is given by the ratio of two quantities: the cardinality of the intersection of the fuzzy sets and the cardinality of the union of the fuzzy sets. The intersection of two fuzzy sets is defined by the minimum operator:

$$(\mu_1 \cap \mu_2)(j) = \min\{\mu_{\mathbf{b}_1}(j)\, \mu_{\mathbf{b}_2}(j)\} \tag{5}$$

The union of two fuzzy sets is defined by the maximum operator:

$$(\mu_1 \cup \mu_2)(j) = \max\{\mu_{\mathbf{b}_1}(j)\, \mu_{\mathbf{b}_2}(j)\} \tag{6}$$

The cardinality of a fuzzy set (also called "$\sigma$-count") is computed by summing up all its membership values:

$$|\mu| = \sum_{j=1}^{m} \mu(j) \tag{7}$$

Summarizing, the similarity between any two users $\mathbf{b}_x$ and $\mathbf{b}_y$ is defined as follows:

$$Sim_{fuzzy}\left(\mathbf{b}_x, \mathbf{b}_y\right) = \frac{\sum_{j=1}^{m} \min\left\{\mu_{\mathbf{b}_x j}, \mu_{\mathbf{b}_y j}\right\}}{\sum_{j=1}^{m} \max\left\{\mu_{\mathbf{b}_x j}, \mu_{\mathbf{b}_y j}\right\}}. \tag{8}$$

This fuzzy similarity measure permits to embed the semantic information incorporated in the user behavior data. In this way, a better estimation of the true similarity degree between two user behaviors is obtained.

Similarity values are mapped into the similarity matrix $\mathbf{Sim} = \left[Sim_{ij}\right]_{i,j=1,\ldots,n}$ where each component $Sim_{ij}$ expresses the similarity value between the user behavior vectors $\mathbf{b}_i$ and $\mathbf{b}_j$ calculated by using the fuzzy similarity measure. Starting from the similarity matrix, the dissimilarity values are simply computed as $Diss_{ij} = 1 - Sim_{ij}$, for $i, j = 1, \ldots, n$. These are mapped in a $n \times n$ matrix $\mathbf{R} = \left[Diss_{ij}\right]_{i,j=1,\ldots,n}$ representing the relation matrix.

## 3 Categorization of User Behaviors

Once the relation matrix has been created, the next activity is the categorization of user behaviors in order to group users with similar interests into a number of user categories. To this aim, we adopt the fuzzy relational clustering approach. In particular, in this work, we employ CARD+, that we proposed in [2] as an improved version of the CARD (Competitive Agglomeration Relational Data) clustering algorithm [8]. A key feature of CARD+ is its ability to automatically categorize the available data into an optimal number of clusters starting from an initial random number. In [8], the authors stated that CARD was able to determine a final partition containing an optimal number of clusters. However, in our experience, CARD resulted very sensitive to the initial number of clusters by often providing different final partitions, thus failing in finding the actual number of clusters buried in data. Indeed, we observed that CARD produces redundant partitions, with clusters having a high overlapping degree (very low inter-cluster distance). CARD+ overcomes this limitation by adding a post-clustering process to the CARD algorithm in order to remove redundant clusters.

As common relational clustering approaches, CARD+ obtains an implicit partition of the object data by deriving the distances from the relational data to a set of $C$ implicit prototypes that summarize the data objects belonging to each cluster in the partition. Specifically, starting from the relation matrix $\mathbf{R}$, the following implicit distances are computed at each iteration step of the algorithm:

$$d_{ci} = \left(\mathbf{R}\mathbf{z}_c\right)_i - \mathbf{z}_c\mathbf{R}\mathbf{z}_c/2 \tag{9}$$

for all behavior vectors $i = 1, \ldots, n$ and for all implicit clusters $c = 1, \ldots, C$, where $\mathbf{z}_c$ is the membership vector for the $c$-th cluster, defined as on the basis of the fuzzy membership values $z_{ci}$ that describe the degree of belongingness of the $i$-th behavior vector in the $c$-th cluster. Once the implicit distance values $d_{ci}$ have been computed, the fuzzy membership values $z_{ci}$ are updated to optimize the clustering criterion, resulting in a new fuzzy partition of behavior vectors. The process is iterated until the membership values stabilize.

Finally, a crisp assignment of behavior vectors to the identified clusters is performed in order to derive a prototype vector for each cluster, representing a user category. Precisely, each behavior vector is crisply assigned to the closest cluster, creating $C$ clusters:

$$\chi_c = \{\mathbf{b}_i \in \mathbf{B} | d_{ci} < d_{ki} \forall c \neq k\} \qquad 1 \leq c \leq C. \tag{10}$$

Then, for each cluster $\chi_c$ a prototype vector $\mathbf{v}_c = (v_{c1}, v_{c2}, \ldots, v_{cm})$ is derived, where

$$v_{cj} = \frac{\sum_{\mathbf{b}_i \in \chi_c} b_{ij}}{|\chi_c|} \qquad j = 1, \ldots, N_P. \tag{11}$$

The values $v_{cj}$ represent the significance (in terms of relevance degree) of a given page $p_j$ to the $c$-th user category.

Summarizing, the CARD+ mines a collection of $C$ clusters from behavior data, representing categories of users that have accessed to the Web site under analysis. Each category prototype $\mathbf{v}_c = (v_{c1}, v_{c2}, ..., v_{cm})$ describes the typical browsing behavior of a group of users with similar interests about the most visited pages of the Web site.

## 4 Simulation Results

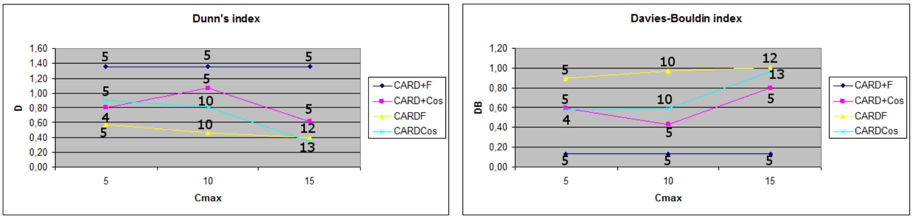
We carried out an experimental simulation to show the suitability of CARD+ equipped with the fuzzy measure to identify Web user categories. We used the access logs from a Web site targeted to young users (average age 12 years old), i.e. the Italian Web site of the Japanese movie Dragon Ball (www.dragonballgt.it). This site was chosen because of its high daily number of accesses (thousands of visits each day).

Firstly, LODAP was used to identify user behavior vectors from the log data collected during a period of 12 hours (from 10:00 a.m. to 22:00 p.m.). Once the four steps of LODAP were executed, a $200 \times 42$ behavior matrix was derived. The 42 pages in the Web site were labeled with a number (see table 1) to facilitate the analysis of results, by specifying the content of the Web pages.

Starting from the available behavior matrix, the relation matrix was created by using the fuzzy similarity measure. Next, the CARD+ algorithm (implemented in the Matlab environment 6.5) was applied to the behavior matrix in order to obtain clusters of users with similar browsing behavior. We carried out several runs by setting a different initial number of clusters $C_{max} = (5, 10, 15)$. To establish the goodness of the derived partitions of behavior vectors, at the end of each run, two indexes were calculated: the Dunn's index and the Davies-Bouldin index [5]. These were used in different works to evaluate the compactness of the partitions obtained by several clustering algorithms. Good partitions correspond to large values of the Dunn's index and low values for the Davies-Bouldin index. We observed that CARD+ with the use of the fuzzy similarity measure provided

**Table 1.** Description of the retained pages in the Web site

| Pages | Content |
|---|---|
| 1, ..., 8 | Pictures of characters |
| 9,..., 13 | Various kind of pictures related to the movie |
| 14,..., 18 | General information about the main character |
| 19, 26, 27 | Matches |
| 20, 21, 36 | Services (registration, login, ...) |
| 22, 23, 24, 25, 28, ..., 31 | General information about the movie |
| 32, ..., 37 | Entertainment (games, videos,...) |
| 38, ..., 42 | Description of characters |



(a)

(b)

**Fig. 1.** Comparison of the Dunn's index (a) and the Davies-Bouldin index (b) obtained by the employed algorithms and similarity measures

data partitions with the same final number of clusters $C = 5$, independently from the initial number of clusters $C_{max}$. The validity indexes took the same values in all runs. In particular, the Dunn's index value was always equal to 1.35 and the value for the Davies-Bouldin index was 0.13. As a consequence, the CARD+ algorithm equipped with the fuzzy similarity measure resulted to be quite stable, by partitioning the available behavior data into 5 clusters corresponding to the identified user categories.

To evaluate the effectiveness of the employed fuzzy similarity measure, we applied CARD+ with the employment of the cosine measure. We carried out the same trials of the previous experiments. Moreover, to establish the suitability of CARD+ for the task of user category identification, we applied the original CARD algorithm to categorize user behaviors by employing either the cosine measure and the fuzzy similarity measure for the computation of the relation matrix. In figure 1, the obtained values for the validity indexes are compared. In this figure, in correspondence of each trial, the final number of clusters extracted by the employed clustering algorithm is also indicated. As it can be observed, CARD+ with the use of the cosine measure derived partitions which categorized data into 4 or 5 clusters, resulting less stable than CARD+ equipped with the fuzzy similarity measure. Moreover, the CARD algorithm showed an instable behavior with both the similarity measures, by providing data partitions with a different final number of clusters in each trial.

Analyzing the results obtained by the different runs, we can conclude that CARD+ with the employment of the fuzzy similarity measure was able to derive the best partition in terms of compactness; hence, it revealed to be a valid approach for the identification of user categories.

The information about the user categories extracted by CARD+ equipped with the fuzzy similarity measure are summarized in table 2. In particular, for each user category (labeled with numbers 1,2,...,5) the pages with the highest degree of interest are indicated. It can be noted that some pages (e.g. $P_1$, $P_2$, $P_3$, $P_{10}$, $P_{11}$, and $P_{12}$) are included in more than one user category, showing how different categories of users may exhibit common interests.

**Table 2.** User categories identified on real-world data

| **User category** | **Relevant pages (interest degrees)** |
|---|---|
| 1 | $P_1(55)$, $P_2(63)$, $P_3(54)$, $P_5(52)$, $P_7(48)$, $P_8(43)$, $P_{14}(66)$, $P_{28}(56)$, $P_{29}(52)$, $P_{30}(37)$ |
| 2 | $P_1(72)$,$P_2(59)$, $P_3(95)$, $P_6(65)$, $P_7(57)$, $P_{10}(74)$, $P_{11}(66)$, $P_{13}(66)$ |
| 3 | $P_1(50)$, $P_2(50)$, $P_3(45)$, $P_4(46)$, $P_5(42)$, $P_6(42)$, $P_8(34)$, $P_9(37)$, $P_{12}(40)$, $P_{15}(41)$, $P_{16}(41)$, $P_{17}(38)$, $P_{18}(37)$, $P_{19}(36)$ |
| 4 | $P_2(49)$, $P_{10}(47)$, $P_{11}(38)$, $P_{12}(36)$, $P_{14}(27)$, $P_{31}(36)$, $P_{32}(29)$, $P_{33}(39)$, $P_{34}(36)$, $P_{35}(26)$, $P_{36}(20)$, $P_{37}(37)$, $P_{38}(29)$, $P_{39}(30)$, $P_{40}(34)$, $P_{41}(28)$, $P_{42}(24)$ |
| 5 | $P_4(70)$, $P_5(65)$, $P_{20}(64)$, $P_{21}(62)$, $P_{22}(54)$, $P_{23}(63)$, $P_{24}(54)$, $P_{25}(41)$, $P_{26}(47)$, $P_{27}(47)$ |

We can give an interpretation of the identified user categories, by individuating the interests of users belonging to each of these. The interpretation is indicated in the following.

- Category 1. Users in this category are mainly interested on information about the movie characters.
- Category 2. Users in this category are interested in the history of the movie and in pictures of movie and characters.
- Category 3. These users are mostly interested to the main character of the movie.
- Category 4. These users prefer pages that link to entertainment objects (games and video).
- Category 5. Users in this category prefer pages containing general information about the movie.

The extracted user categories may be used to implement personalization functions in the considered Web site.

## 5 Conclusions

A fuzzy clustering approach has been presented to identify user categories starting from Web usage data. A fuzzy measure has been proposed to evaluate similarity between Web users. The measure has been integrated in CARD+, a relational fuzzy clustering algorithm, in order to capture similarity in usage data and derive user categories. Comparative results have shown that CARD+ equipped with the fuzzy similarity measure overcomes CARD+ equipped with the standard cosine similarity measure. Also, it overcomes the original CARD algorithm, whatever the adopted measure is. Clusters derived by CARD+ using the fuzzy measure are sufficiently separate and correspond to actual user categories embedded in the available log data. The identified user categories will be exploited to realize personalization functionalities in the considered Web site, such as the dynamical suggestion of links to pages considered interesting for a current user, according to his category membership.

## References

1. Bezdek, J.C.: Pattern recognition with fuzzy objective function algorithms. Plenum Press, New York (1981)
2. Castellano, G., Fanelli, A.M., Torsello, M.A.: Relational Fuzzy approach for Mining User Profiles. Lectures Notes in Computational Intelligence, pp. 175–179. WSEAS Press (2007)
3. Castellano, G., Fanelli, A.M., Torsello, M.A.: LODAP: A Log Data Preprocessor for mining Web browsing patterns. In: Proc. of The 6th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases (AIKED 2007), Corfu Island, Greece (2007)
4. Facca, F.M., Lanzi, P.L.: Mining interesting knowledge from weblogs: a survey. Data and Knowledge Engineering 53, 225–241 (2005)
5. Halkidi, M., Batistakis, Y., Vazirgiannis, M.: Cluster Validity Methods: Part II. SIGMOD Record (2002)
6. Joshi, A., Joshi, K.: On mining Web access logs. In: ACM SIGMOID Workshop on Research issues in Data Mining and Knowledge discovery, pp. 63–69 (2000)
7. Mobasher, B., Cooley, R., Srivastava, J.: Automatic personalization based on Web usage mining. TR-99010, Department of Computer Science. DePaul University (1999)
8. Nasraoui, O., Frigui, H., Joshi, A., Krishnapuram, R.: Mining Web access log using relational competitive fuzzy clustering. In: Proc. of the Eight International Fuzzy System Association World Congress (1999)
9. Suryavanshi, B.S., Shiri, N., Mudur, S.P.: An efficient technique for mining usage profiles using Relational Fuzzy Subtractive Clustering. In: Proc. of WIRI 2005, Tokyo, Japan (2005)
10. Vakali, A., Pokorny, J., Dalamagas, T.: An Overview of Web Data Clustering Practices. In: Lindner, W., Mesiti, M., Türker, C., Tzitzikas, Y., Vakali, A.I. (eds.) EDBT 2004. LNCS, vol. 3268, pp. 597–606. Springer, Heidelberg (2004)
11. Wang, X., Abraham, A., Smith, K.A.: Intelligent web traffic mining and analysis. Journal of Network and Computer Applications 28, 147–165 (2005)
12. Zhizhen, L., Pengfei, S.: Similarity measures on intuitionistic fuzzy sets. Pattern Recognition Letter 24, 2687–2693 (2003)

# Fuzzy User Profiling in e-Learning Contexts

Corrado Mencar, Ciro Castiello, and Anna Maria Fanelli

Department of Informatics – University of Bari
Via E. Orabona, 4 – 70125 Bari Italy
{mencar,castiello,fanelli}@di.uniba.it

**Abstract.** The research activity described in this paper concerns the personalisation process in e-learning contexts. Particular emphasis is laid on the mechanisms of user profiling and association between user profiles and pedagogical resources. A particular profiling model is proposed where both the pedagogical resources and the user profiles are described in terms of a fuzzy valued metadata specification. The adoption of specific fuzzy operators enables the proposed model to perform associations with a high degree of flexibility, yielding a customised resource allocation for each user.

## 1 Introduction

The last decades have witnessed a growing interest for adaptive software systems, which are able to take into account the peculiarities of the distinct users in order to improve the interaction possibilities. The demand for adaptive systems directly follows the increasing development of Web applications: the huge number of users connecting on-line requires suitable tools for the personalisation of Web contents [1]. The basic idea consists in realising interactive systems with the capability of assigning to each user the contents which best match his interests and preferences. Such a personalisation process is basically founded on two steps:

1. the automatic construction of models (user profiles) which represent the user characteristics, interests and preferences;
2. the automatic selection of the contents to be proposed, on the basis of the previously identified user models.

The efficacy of a personalisation process, therefore, is strictly connected with the possibility of an automatic detection of the user profiles, through the analysis of the users' behaviour during their interactions with the system.

Personalisation processes have been successfully adopted in several real-world applications, with the realisation of e-commerce infrastructures, information retrieval systems, digital libraries, and so on [2]. Moreover, personalisation processes have been applied in the context of e-learning [3], with the aim of identifying educational courses tailored on the single user's requirements. In this way, it is possible the definition of e-learning techniques as «user-centred» teaching solutions. Starting from the early forms Adaptive Learning Environments [4], the

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 230–237, 2008.

research in this field led to the development of the current Personal Learning Environments [5] which constitute the most advanced versions of e-learning systems providing tools for configuring and managing the user learning experiences.

The personalisation process requires the contemporary definition of both user and resource models: they should be compatible in order to determine the proper association between user profiles and pedagogical contents. This paper proposes a profiling system where both the pedagogical resources and the user profiles are described by means of metadata, with values represented by fuzzy sets. In this way, the association between profiles and resources is realised in terms of fuzzy compatibility degrees. The employment of fuzzy sets and fuzzy operators allows for non-binary associations and a compatibility ranking can be arranged, so that it is possible to distinguish among resources with a greater or a lower compatibility degree with respect to a specific user.

The paper is organised as follows. In the next section the model of the proposed profiling system is described. Section 3 introduces the scheme of the profiling system in the context of an e-learning application. Section 4 closes the paper with some conclusive remarks.

## 2 The Model of the Profiling System

The model of the proposed profiling system has been conceived in order to satisfy the following requirements:

- independence from the pedagogical resource representation and the user description;
- possibility of describing complex profiles, to which the users may be associated even partially;
- possibility of managing imprecise descriptions both of the pedagogical resources and user profiles.

By satisfying those principles, it is possible to bring forward a highly modular development of the profiling system which can be only weakly connected with other components inside an e-learning environment.

### 2.1 Resourse Description

Each pedagogical resource is formally represented by metada describing its semantics. Inside the profiling system, a resource is characterised only in terms of its description, that is the ensemble of the associated metadata:

$$DESCR(resource) = \{metadata\},$$

where each metadata associates a resource with an Attribute and a Value:

$$metadata = (Attribute, Value).$$

Actually, some metadata can not be precisely represented by single or collective values. Some attributes, in fact, are characterised by imprecise values: that is the case of the «fruition time» (short, long, etc.), the «learning complexity» (easy, average, expert, etc.), and so on. The classical systems for metadata definition rely on discretisation of the attribute domain. This kind of approach is based on arbitrary choices and may produce unexpected results. In this work, the employment of fuzzy logic is advocated to represent different properties (which can be precise or imprecise) in a homogeneous form. Fuzzy logic has been proposed by Zadeh [6] as an extension of classical logic and it is based on the concept of fuzzy set, which differentiates from the conventional set by introducing the degree of membership of an object with respect to a set. Precisely, a fuzzy set is defined by a membership function associating each domain element with a membership degree, that is a value in the range $[0, 1]$:

$$fuzzy\_set : domain \rightarrow [0, 1].$$

By applying the concept of fuzzy set to resource representation, the attribute value for a specific resource is a fuzzy set defined over the attribute domain (instead of a single element inside the attribute domain):

$$attribute\_value : attribute\_domain \rightarrow [0, 1].$$

The adoption of fuzzy sets allows the uniform modelling of diverse attributes:

- punctual value attributes
  (for instance: dimension= {“300Kb”/1});
- collective value attributes
  (for instance: scope= {“Computer Science”/1,“Administration”/0.5});
- imprecise value attributes
  (for instance: fruition time=short, being «short» a fuzzy set defined over temporal quantities).

Punctual and collective attributes can be described by means of an intensional fuzzy metadata process, with fuzzy sets represented in terms of pair sequences (value/membership degree). Imprecise attributes can be described by means of an extensional fuzzy metadata process, with fuzzy sets defined over the continuous attribute domain (trapezoidal fuzzy sets offer a compromise between simplicity and generality).

The description of an illustrative pedagogical resource is reported in table 1, where the fuzzy Attribute-Value pairs are reported. As it can be observed from the analysis of the table, the «Title» attribute is punctually valorised by means of the resource title. The same goes with the «Topic» attribute. The «Scope» attribute assumes two distinct values with different membership degrees. This kind of granular valorisation is adopted also for the «Keywords» attributes (it can be noted that the keywords related to the resource are distinct from those related to the prerequisites to access the resource). Finally, the «Fruition time» attribute is described in terms of a trapezoidal fuzzy set, with the indication of its associated parameters. In this way, the fruition time is represented in an imprecise form, that is a time range of about 30-60 minutes, with a minimum

**Table 1.** Description of a pedagogical resource by means of fuzzy metadata

| **Attribute** | **Fuzzy Value** |
|---|---|
| Title | {"Basic Word Course"/1.0} |
| Topic | {"Computer Science"/1.0} |
| Scope | {"Word Processing"/1.0,"File management"/0.2} |
| Keywords<br>(related to the resource) | {"Word Processing"/1.0,"File management"/0.2,<br>"Word"/1.0,"Windows"/0.2} |
| Keywords<br>(prerequisite) | {"Windows"/0.7,"File management"/0.5} |
| Fruition time | $T(15, 30, 60, 90)$ |

and a maximum time equal to 15 and 90 minutes, respectively. The illustrative metadata process is not intended as an exhaustive valorisation including all the possible attributes associable to the resource. In fact, a thorough valorisation is not mandatory for the application of the proposed profiling model (even if some attributes, such as the title, appear to be of primary relevance).

### 2.2 User Profile Description

The user profiles are data structures characterising user stereotypes. In order to take into account stereotypes however complex, the structure of the user profile is defined in terms of an ensemble of profile components:

$$profile = \{profile\_components\}.$$

The profile components represent basic profiles and they are described in terms of metadata specifications which are analogous to those adopted for the resource description, that is:

$$profile\_component = \{metadata\}.$$

The description of an illustrative profile component is reported in table 2. The example refers to a manager profile, with interest to word processing and spreadsheet utilities. To a lesser extent, this kind of user is interested also to database instruments. Moreover, the profile is characterised also by a good acquaintance with the Windows operating system and a quite good acquaintance with file management. The fruition time to be devoted to the pedagogical resources is about 15-30 minutes, with a maximum time equal to 60 minutes.

In real world situations it is quite uncommon that a user may be characterised by a single profile. Moreover, the membership to different profiles can be partial. To improve the adaptability of the profiling system, the proposed approach is based on a (possibly) partial association of a user to different profiles:

$$DESCR(user) : \{profiles\} \rightarrow [0, 1].$$

Three categories of profiles have been identified to be used inside the proposed profiling system:

**Table 2.** Description of a profile component by means of fuzzy metadata

| **Attribute** | **Fuzzy Value** |
|---|---|
| Role | {"Manager"/1.0} |
| Keywords<br>(related to the user) | {"Word Processing"/1.0,"Spreadsheet"/1,<br>"Database"/0.3} |
| Keywords<br>(prerequisite) | {"Windows"/1,"File management"/0.8} |
| Fruition time | $T(0, 15, 30, 60)$ |

1. expertise profiles, characterising the role of the users inside the e-learning environment. The expertise profiles are shared among users and can be used to suggest the association of the pedagogical resources;
2. preference profiles, characterising the preferences of users, not to be shared among them. The preference profiles can be used to suggest the association of the pedagogical resources;
3. acquaintance profiles, characterising the users in terms of the specific information they possess. The acquaintance profiles are not to be shared among the users and can be used to filter all the pedagogical resources which, although compatible with the other profile categories, have already been associated to the users in the past.

For each user, three profile bases are identified (corresponding to the previously described categories):

$$\begin{gathered} Expertise = \{profiles\} \\ Preference(user) = \{profiles\} \\ Acquaintance(user) = \{profiles\}. \end{gathered}$$

It can be noted how it is not necessary to specify the user for the expertise profile base, since it is shared among all the users.

### 2.3 Associating User Profiles with Resources

The homogeneous representation adopted allows the association of the pedagogical resources with the profile components on the basis of a matching process: the corresponding metadata can be compared by applying fuzzy operators. Actually, it is possible to define a compatibility measure among metadata in terms of the possibility measure among fuzzy sets:

$$COMP(metadata_1, metadata_2) = POSS(fuzzy\_set_1, fuzzy\_set_2),$$

where

$$POSS(value_1, value_2) = MAX_x(MIN(\mu_{value1}(x), \mu_{value2}(x))),$$

being $\mu_{fuzzy\_set}(x)$ the membership degree of the element $x$ inside the attribute domain to the fuzzy set representing the attribute value.

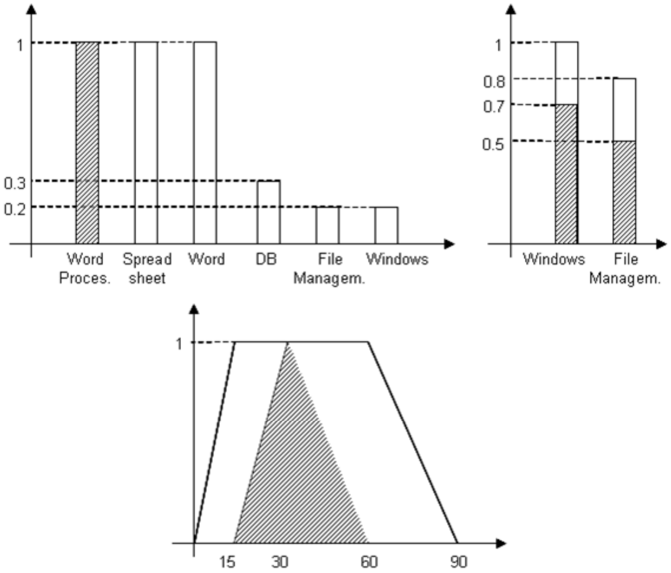


**Fig. 1.** Compatibility between a pedagogical resource and a component profile

The compatibility degree between a pedagogical resource and a profile component can be evaluated in terms of aggregation of the previously computed compatibility degrees:

$$COMP(resource, profile\_component) = AGGR\{COMP(m_r, m_c)\},$$

being $m_r$ a resource metadata and $m_c$ a profile component metadata, with the constraint that both metadata are related to the same attribute. The aggregation operator is defined by means of a parameterised mean.

As an example, the evaluation of the compatibility degree between the resource described in table 1 and the profile component described in table 2 is graphically represented in figure 1. It can be noted how the compatibility is evaluated on the basis of the common set of attributes (that are: «Keywords» and «Fruition time»). The aggregation operator can be applied over the obtained compatibility degrees (highlighted in figure): by adopting the minimum function, the final compatibility value is equal to 0.7.

The compatibility degree between a resource and a user profile is obtained by aggregating the compatibility degrees between the resources and all the components of the profile. In this case, the aggregation operator is defined as the maximum compatibility degree:

$$COMP(resource, profile) = MAX\{COMP(resource, profile\_component)\}.$$

Finally, the compatibility degree between a resource and a profile base is defined as the maximum compatibility degree between the resource and all the profiles inside the base. The evaluation must take into account the partial membership of the user to each profile:

$$COMP(resource, base, user) = MAX \\ \{MIN(COMP(resource, profile), \mu_{user}(profile)) | profile \in base\}.$$

The specific meaning of the different profile bases which have been previously introduced plays a role in specifying the compatibility between the given resource and the single user, which is defined as follows:

$$COMP(resource, user) = MIN(MAX(COMP(resource, Expertise, user), \\ COMP(resource, Preference, user)), \\ 1 - COMP(resource, Acquaintance, user)).$$

## 3 The Profiling System Inside an e-Learning Context

The model of the profiling system described in the previous section may find application inside an e-learning context, with the aim of providing an automatic allocation of resources to different categories of users. The proposed model may represent the basis for one of the modules composing the overall architecture of the e-learning environment. Other modules may be devoted to the management of the pedagogical contents and the organisation of the necessary databases for the objects involved in the environment. In this section, the profiling module is illustrated, taking for granted the independence requirements with respect to all the other possibly involved modules.

The general scheme of the profiling module is depicted in figure 2. Two databases are included in the scheme: the Resource Description (R-D) and the Profile Description (P-D) databases, which are related to the descriptions of the pedagogical contents and of the user profiles, respectively. The representation of resources and profiles by means of their descriptions (which are in both cases

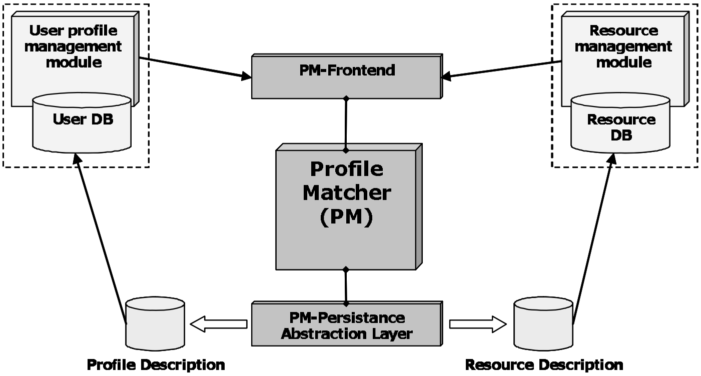

User profile management module
User DB
PM-Frontend
Resource management module
Resource DB
Profile Matcher (PM)
PM-Persistance Abstraction Layer
Profile Description
Resource Description

**Fig. 2.** The scheme of the profiling module and its connections with other management modules

compiled in terms of the metadata specifications illustrated in the previous section) is due to the requirement of independence from the specific representations adopted outside of the profiling module.

Among the software components of the profiling module, the Profile Matcher component (PM) represents the core of the profiling mechanism. The PM, in fact, is responsible for the automatic association process involving resources and user profiles, by means of the compatibility evaluation based on fuzzy operators. In practice, the PM provides the index of the pedagogical contents to be addressed to a particular user profile, ranked in terms of compatibility degrees. The communication tasks of the PM are executed by a couple of interface components: the PM-Persistence-Abstraction-Layer (PM-PAL) and the PM-Frontend (PM-FE). The PM-PAL realises the interface between the PM and the R-D and P-D databases. The employment of the PM-PAL is due to the requirement of independence from the actual persistence layer. The PM-FE realises the interface between the PM and the management modules which are part of the overall e-learning environment architecture.

## 4 Conclusions

A profiling model which makes use of a fuzzy metadata specification is presented in this paper. The proposed approach allows a suitable representation of pedagogical resources and user profiles, together with a more comprehensive association among them, which is realised in terms of fuzzy compatibility degrees. The model can find application in e-learning environments and the scheme of a specific profiling module has been illustrated, addressing also the mechanisms of integration inside a more general architecture of an e-learning platform.

## References

1. Nasraoui, O.: World wide web personalization. In: Wang, J. (ed.) Encyclopedia of Data Mining and Data Warehousing. Idea Group (2005)
2. Tasso, C., Omero, P.: La personalizzazione dei Contenuti Web: E-Commerce, I-Access, E-Government. Franco Angeli (2002)
3. Dolog, P., Henze, N., Nejdl, W., Sintek, M.: Personalization in distributed elearning environments. In: Proceedings of WWW 2004 - The Thirteen International World Wide Web Conference (2004)
4. Brusilowsky, P.: Adaptive and intelligent technologies for web-based education. Intelligent Systems and Teleteaching 4, 19–25 (1999)
5. Adler, C., Rae, S.: Personalized learning environments: The future of e-learning is learner-centric. E-learning 3(1), 22–24 (2002)
6. Zadeh, L.: Fuzzy sets. Information and Control 8, 338–353 (1965)

# Personalized Web Search by Constructing Semantic Clusters of User Profiles

John Garofalakis[1,2], Theodoula Giannakoudi[1,2], and Agoritsa Vopi[1,2]

[1] RA Computer Technology Institute
Telematics Center Department
N. Kazantzaki str. 26500, Greece
[2] University of Patras
Computer Engineering & Informatics Dept
26500 Patras, Greece
garofala@cti.gr, {gianakot, vopi}@westgate.gr

**Abstract.** During the recent years the Web has been developed rapidly making the efficient searching of information difficult and time-consuming. In this work, we propose a web search personalization methodology by coupling data mining techniques with the underlying semantics of the web content. To this purpose, we exploit reference ontologies that emerge from web catalogs (such as ODP), which can scale to the growth of the web. Our methodology uses ontologies to provide the semantic profiling of users' interests based on the implicit logging of their behavior and the on-the-fly semantic analysis and annotation of the web results summaries.

**Keywords:** Web Usage Mining, Semantic Annotation, Clustering, Ontology, User Profiles, Web Search, Personalization.

## 1 Introduction

While Web is constantly growing, web search is becoming more and more a complex and confusing task for the web user. The vital question is which the right information for a specific user is and how this information could be efficiently delivered, saving the web user from consecutive submitted queries and time-consuming navigation through numerous web results.

Most existing Web search engines return a list of results based on the query without paying any attention to the underlying user's interests or even to the searching behaviors of other users with common interests. There is no prediction of the user's information needs and problems of polysemy and synonymy often arise. Thus, when a user submits searching keywords with multiple meaning (polysemy) or several words having the same meaning with the submitted keyword (synonymy), he will probably get a large number of web results and most of them will not meet his need. For, example, a user submitting the term "opera" may be interested in arts or computers but the results will be the same regardless of what he looks for.

Some current search engines such as *Google* or *Yahoo!* have hierarchies of categories to provide users with the opportunity to explicitly specify their interests. However,

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 238–247, 2008.

these hierarchies are usually very large; therefore, they discourage the user from browsing them in order to define the interested paths. To overcome these overloads in the users searching tasks, the user interests may be implicitly detected by tracking his search history and personalizing the web results.

In this work, we propose a personalization method, which couples data mining techniques with the underlying semantics of the web content in order to build semantic clusters of user profiles. Regarding the *semantic clusters*, they actually comprise taxonomical subsets of a general category hierarchy, such as ODP-Open Directory Project [17], representing the categories of interest for groups of web users with similar search tasks. In out methodology, apart from exploiting a specific user search history, we further exploit the search history of other users with similar interests. The user is assigned to relevant classes of common interest, so as to predict the relevance score of the results with the user goal and finally re-rank them. To this purpose, we exploit reference ontologies that emerge from web catalogs (such as ODP), which can scale to the growth of the web.

Specifically, our methodology consists of five tasks: (1) gathers user's search history, (2) processes the user activity, taking into consideration other users' activities and constructing clusters of commonly preferred concepts, (3) defines ontology-based profiles for the active user based on the detected interests from his current activity and the interests depicted from the semantic cluster in which he has been assigned from previous searching sessions, (4) re-ranks the web results combining the above information with the semantics of the delivered results and (5) constantly re-organizes the conceptual clusters in order to be up-to-date with the users' interests.

Our approach has been experimentally evaluated by utilizing the Google Web Service and the results show that semantically clustering users in terms of detecting commonly interesting ODP categories in search engines is effective.

The remainder of the paper is structured as follows: Section 2 discusses related work. In Section 3, we describe the ODP-based reference ontology that our approach uses. Using this ontology, we outline the semantic annotation of web results to the ontology classes. Moreover, we present how the user profiles are defined over the reference ontology referred earlier as task (2) and how the semantic user clusters are formed, referred as task (3). In Section 4, we propose a novel technique for web search personalization combining profiles of semantic clusters with the emerging profile of the active user referred as tasks (4) and (5). In Section 5, we exhibit our experiments. Section 6 presents the conclusions and gives an outlook on further work.

## 2 Related Work

In this section, we present work that has been conducted in similar contexts, such as personalized web searching, usage-based personalization and semantic-aware personalization.

Several ontology-based approaches have been proposed for users profiling taking advantage of the knowledge contained in ontologies ([4], [9]) in personalization systems. In [3], an aggregation scheme towards more general concepts is presented. Clustering of the user sessions is provided to identify related concepts at different levels of abstraction in a recommender system.

Significant studies have been conducted for personalization based on user search history. A general framework for personalization based on aggregate usage profiles is presented in [13]. This work distinguishes between the offline tasks of data preparation and usage mining and the online personalization components. [15] suggests learning a user's preferences automatically based on their past click history and shows how to use this learning for result personalization.

Many researchers have proposed several ways to personalize web search through biasing ranking algorithms towards possible interesting pages for the user. [16] extends the HITS algorithm to promotes pages marked "relevant" by the user in previous searches. A great step towards biased ranking is performed in [6], where a topic-oriented PageRank is built, considering the first-level topics listed in the Open Directory. The authors show this algorithm overperforms the standard PageRank if the search engine can effectively estimate the query topic.

The authors show this algorithm overperforms the standard PageRank if the search engine can effectively estimate the query topic.

Specifically, regarding the exploitation of large-scale taxonomies in personalized search, a number of interesting works has been presented. In [2], several ways are explored of extending ODP metadata to personalized search. In [8], users' browsing history is exploited to construct a smaller subset of categories than the entire ODP and a novel ranking logic is implemented. In [7], sets of known user interests are automatically mapped onto a group of categories in ODP and manually edited data of ODP are used for training text classifiers to perform search results categorization and personalization.

Our work differs from previous works in several tasks. We exploit large-scale taxonomies, such as ODP, to construct combinative semantic user profiles. In our emerging profiles, both user browsing history and automatically created clusters of user categories are incorporated in personalizing web results. In this way, we re-rank search results taking under consideration apart from the active user tasks, the subsets of "interesting" taxonomy categories that co-occur in other users searches, in the case that these users exhibit similar behavior with the active one.

## 3 Ontology-Based User Clusters

The general aim of this work is to introduce a method for personalizing the results of web searching. For this reason, we focused on constructing user profiles implicitly and automatically, according to their interests and their previous behavior on searching. At this direction we were based on the work described in [1].

### 3.1 Reference Ontology

Our first goal was to create a reference ontology for the user profiles. The profile of each user will be represented by a weighted ontology, depicting the users' interest for every class of the reference ontology. We implemented an OWL ontology based on ODP [17] using Protégé [18].

The ontology created is actually a directed acyclic graph (DAG). Since we wish to create a relatively concise user profile that identifies the general areas of a user's interests we created our reference ontology by using concepts from only the first three levels of ODP, which are the directories used by Google search Engine. In addition, since we want concepts that are related by a generalization-specialization relationship, we remove subjects that were linked based on other criteria, e.g. alphabetic or geographic associations.

### 3.2 Semantic Annotation

The construction of the profile, i.e. the weighted ontology, for every user includes the semantic annotation of the user's previous choices. The semantic characterization of the user choices is the one proposed in [5]. The user's previous choices are analyzed into keywords and the keywords are semantically characterized. The calculation of the semantic similarity between each keyword and each term of the ontology was computed by using semantic similarity measures with WordNet [10][11]. The measure that was applied in our methodology is the Wu & Palmer [12] one. This measure calculates the relatedness by considering the depths of the two synstes (on or more sets of synonyms) in the WordNet taxonomies, along with the depth of the LCS (Lexical Conceptual Structure).

$$\text{Score} = \frac{2 * depth(lcs)}{(depth(s1) + depth(s2))}$$

This means: score$\in$ (0,1]. The score can never be zero because the depth of the LCS is never zero. The score is one if the two input synsets are the exactly same.

After applying semantic annotation, the keywords and consequently, the users' choices are assigned to relevant classes of the ontology after the completion of the ontology assignment step in the proposed method.

### 3.3 Definition of User Profiles

In this step, the semantic annotations of the users' choices are utilized to construct the profile for every user. The method applied is similar to [13]. From the web access logs kept in the web server our method extracts the user's previous choices, which have already been semantically annotated. Therefore, for every user, we extract the concepts and the frequency of appearance from the previous choices that the specific user has made. In the end of this step, there is an accumulation of the preferences for every user and of the frequency for every concept, which is the weight, for every class (preference) in the ontology.

In this step, apart from the accumulation of the concepts for which the user has shown interest, we construct the vector that represents each user's profile. The vector's size is the number of concepts that the ontology consists of. The value of each element of the vector corresponds to the weight of the user interest for this concept.

The weight for a concept i for the user u, is calculated as:

$$w_{iu} = \frac{cf_{iu}}{sum(cf_u)} ,$$

where $cf_{iu}$ = the number of times that the concept i has been assigned to the user u.

$cf_u$ = the sum of the times that all the concepts of the ontology has been assigned to the user u.

For the concepts that there is no previous assignment of the user the value is set to zero.

So for a user the profile is represented as follows:

$$p = < w_1^p, w_2^p, ..., w_n^p > \tag{1}$$

Where n is the number of concepts in the ontology and

$$w_i^p = \{ \begin{matrix} weight(concept_i, p) if concept_i > 0 \\ 0, otherwise \end{matrix}$$

Therefore, the weight of each concept is the relative frequency of the concept among all concepts of the ontology. The sum of all weight is equal to one, representing the percentage of the user's interest for every concept. Moreover, for each user we create a file that has the profile vector.

### 3.4 Semantic Clustering of User Profiles

After creating each user profile, the suggested methodology moves on profile clustering (figure 1). From the profile creation step, a profile for every user is stored in the database and a file with the user's vector weighted ontology is created. At this step, the profiles of all the users that reacted with the search engine are accumulated and are grouped into clusters with similar interests.

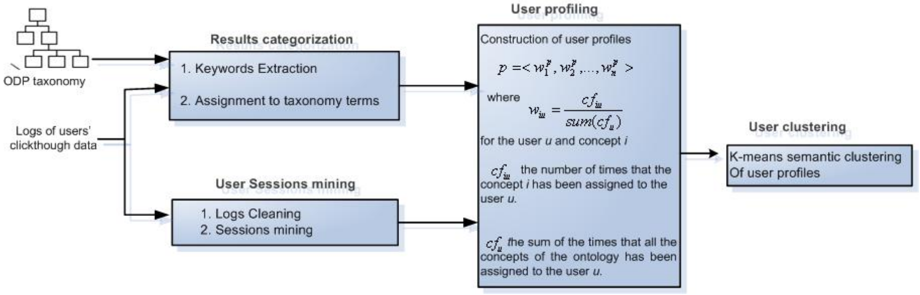

ODP taxonomy
Logs of users' clickthough data
Results categorization
1. Keywords Extraction
2. Assignment to taxonomy terms
User Sessions mining
1. Logs Cleaning
2. Sessions mining
User profiling
Construction of user profiles
$p = < w_1^p, w_2^p, ..., w_n^p >$
where
$w_{iu} = \frac{cf_{iu}}{sum(cf_u)}$
for the user u and concept i
$cf_{iu}$ the number of times that the concept i has been assigned to the user u.
$cf_u$ the sum of the times that all the concepts of the ontology has been assigned to the user u.
User clustering
K-means semantic clustering Of user profiles

**Fig. 1.** Creation of the Semantic Users' Profiles

The clustering algorithm that has been applied is the K-Means [14]. K-Means is one of the most common clustering algorithms that groups data into clusters with similar characteristics together. In the end of the execution of this step the users are grouped into clusters with similar interests and the clusters are stored to the database. We should note that every time this step is executed, the clusters are constructed from the beginning and the users are grouped again. Thus, the clustering procedure is not based on the previous constructed clusters. This has been chosen as a way of developing the methodology, considering that the user's choices will alter periodically.

## 4 Personalization Algorithm

The preprocessed user's choices, their semantic characterization and the users' clusters are used for processing and personalizing the web search results. At this point every user that has reacted previously with the search engine has been assigned in one cluster. Every cluster consists of users with similar interests and can be depicted as a weighted ontology. The elements of the vector, representing the weighted ontology, would be the sum of interests for a concept of all the users belonging to the cluster divided by the sum of interests of all the users of the cluster for all the concepts of the ontology. The formulation is the same that was followed in the users' profiles described in paragraph 3.3.

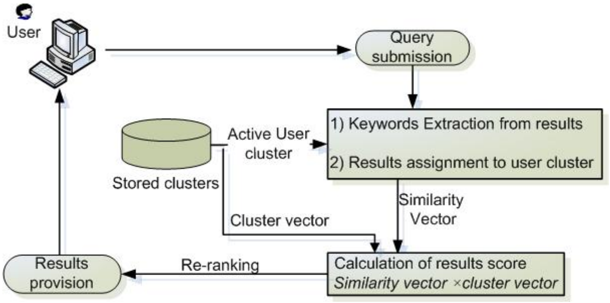

User
Query submission
Active User cluster
1) Keywords Extraction from results
2) Results assignment to user cluster
Stored clusters
Similarity Vector
Cluster vector
Results provision
Re-ranking
Calculation of results score
Similarity vector ×cluster vector

**Fig. 2.** The Personalization algorithm

The personalized search includes the calculation of the similarity of each search result with the cluster's interests. This calculation requires the execution of all the steps of the ontology-based user clusters for each result. Therefore, for every query the following steps are performed:

1) Extracts the keywords from the results returned
2) Applies the semantic annotation step with the difference that at this assignment the ontology is not the whole reference ontology but a part of it which consists of the concepts of the ontology for which the cluster that the user belongs has a non-zero weight. The output of this step is a vector containing the similarity values of keywords with the concepts of the ontology and is depicted as:

$$result\_sim_{jc} = < sim_1^j, sim_2^j, ..., sim_m^j > (2)$$

Where: $j$ is the jth result of the search engine and $m$ is the number of the concepts in the cluster.

Also, each element of this vector is depicted as follows:

$$sim^j = \max(sim_1^k, sim_2^k, ..., sim_m^k)$$

Where: $k$ is the number of the keywords and $m$ is the number of the concepts in cluster.

3) Since we have calculated the similarity of each result to the cluster we calculate the score value for each result. This score is calculated as the internal

product of the cluster vector represented in relation (1) and the similarity vector represented in relation (2).

So the score will be:

$$Score = c \times result\_sim.$$

The above three steps are executed for every result and the score value is kept in cache. Afterwards, the results of the search engine are organized for presentation to the user according to the score that has been calculated, beginning with the one with the highest score (Figure 2).

## 5 Testing and Evaluation

In order to evaluate the proposed method and prove the efficient behavior of our personalization method, we performed some queries with polysemy expecting the personalized results to be personalized according to the profile of the cluster that a user is set. We applied the queries in an experimental implementation that utilizes the Google search API. In one case, we applied our personalization methodology, whereas in the other case we extracted the results as they were returned by the search API. In this paper, we present one of the queries.

The Google search API, used for the experimental implementation, returns the URL, the title and a summary for every result. We used a database for storing the users' information and choices for every query submitted. Through the website we stored the IP address, the domain name and the user agent for the identification of each user. Moreover, the search engine stores in the database the query and the choices of the user for every query. So, for every result that is clicked by the user the search engine stores the title, the URL and the short summary returned in the database. This database consists of the history of the requests and therefore is used as the web access logs in this methodology. At next, we apply the steps of the methodology proposed earlier in the web access logs for the creation of the semantic users' profiles clusters.

Our experimental implementation was tested for two weeks by twenty users. The choices that they have made for every query were stored in the database. They were processed and the users' profiles were created. Next, we clustered the users in three clusters. The user that made the queries has already been put in a cluster and the reference ontology of the cluster upon which the score of the results will be based has been created. We should note that the cluster has users that are interested in Acting, Advertising, American, Animation, Apple, Appliances, Artists, Audio, Ballet, Ballroom, Biography, Bonsai, Buses, Cables, Choices, Companies, Darwin, DEC, Exploits, Flowers, Fraud, Games, Journals, Licenses, Mach, Mainframe, Morris, Mosaics, Music, Oceania, Opera, Painters, People, Pick, Programs, Quotations, Reference, Representatives, Roleplaying, Security, Series, Soaps, Sports, Sun, Supplies, Syllable, Telephony, Test Equipment, Youth, Assemblage, Characters, Christian, Computer, Cracking, Creativity, Creators, Drawing, Editorial, Home, Instruments, Internet, Organizations, Radio, Searching, Unix with various weights for each concept of the reference ontology.

The query we applied in the search engine was "opera". The word "opera" has a twofold meaning. Opera is a form of musical and dramatic work and also it is a very common used web browser. Thus, it is a query that the results of the search engines will refer both to music and computers. The user that is giving the query to the search engine asks for information about opera as a kind of music and expects results related to music. In table 1, the results of the search engine are presented. The first column represents the order of the results of the search API without any personalization, while in the second column the order of our personalized result is presented. Next to each title we give in parenthesis the general concept of the result.

The first column represents the results that are returned from the search API without personalization. In this column the results that the user searches are in places 3, 6, 8, 9. On the other hand the second column has the personalized results and the results related with music are in places 2, 3, 6, 8. It is obvious that in the personalized results, the pages related with music are ranked higher. The cluster into which the user belongs includes many music topics and this has been taken into consideration while calculating the score of each result pushing the results related with music in higher positions. Also, because of the fact that the results returned have high similarity with the concepts of the cluster reference ontology the music related results are pushed closer to the top. In our example, the cluster that the user belongs except for the interest in music shows also interest in computers, and this interest is depicted in the results of the personalization methodology applied. The first result in both queries was about computers because the weighted ontology depicting the cluster has higher weights for concepts related to computers than concepts related to arts. However, the methodology given the relatedness of the results with the cluster's preferences has pushed the desired results in places higher than the places they were put without personalization.

**Table 1.** Personalized and non-personalized results for query "Opera" for a user interested in opera related with music, while the cluster he belongs has interest in Arts and Computes

| Non Personalized Results | Personalized Results |
|---|---|
| Download Opera Web Browser (computers) | Opera Software-Company (computers) |
| Opera Software-Company (computers) | Welcome to LA Opera I LA Opera (music) |
| Opera - Wikipedia the free encyclopedia (music) | Opera - Wikipedia the free encyclopedia (music) |
| Opera (Internet suite) – Wikipedia, the free encyclopedia (computers) | Opera Community (computers) |
| Opera Mini – Free mobile Web browser for your phone (computers) | Opera (Internet suite) – Wikipedia, the free encyclopedia (computers) |
| Welcome to LA Opera I LA Opera (music) | Opera in to the Ozarks (music) |
| OperaGlass (computers) | Opera Mini – Free mobile Web browser for your phone (computers) |
| The Metropolitan Opera (music) | The Metropolitan Opera (music) |
| Opera in to the Ozarks (music) | OperaGlass (computers) |
| Opera Community (computers) | Download Opera Web Browser (computers) |

## 6 Conclusions and Future Work

In this paper, we presented a personalization methodology which is based on clustering semantic user profiles. The method analyzes and annotates semantically the web access logs. Next, it organizes the users' profiles and groups the users into clusters. The personalization of the web search results is performed through an on-the-fly semantic characterization and the score of each result is calculated. The scores of the results are kept in cache and the results are reorganized and presented to the user according to this score putting the one with the highest score first. By the experimental implementation we showed that the personalized method proposed is effective. Future work includes the use of Fuzzy K-Means that allows the creation of overlapping clusters, so that a user may belong to different cluster profiles with different weights. Also, the development of a reference ontology with more levels and alteration in factors such as the score of each result taking into consideration the user's preference with greater weight than the rest users of the cluster.

## References

1. Cauch, S., Chafee, J., Pretschner, A.: Ontology-Based User Profiles for Search and Browsing. Web Intelligence and Agent systems 1(3-4), 219–234 (2003)
2. Chirita, P.A., Nejdl, W., Paiu, R., Kohlschütter, C.: Using ODP metadata to personalize search. In: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, Salvador, Brazil (August 2005)
3. Dai, H., Mobasher, B.: Using Ontologies to Discover Domain-Level Web Usage Profiles. In: Proc. of the 2nd Workshop on Semantic Web Mining. PKDD 2002, Helsinki, Finland (August 2002)
4. Eirinaki, M., Vazirgiannis, M., Varlamis, I.: SEWeP: Using Site Semantics and a Taxonomy to Enhance the Web Personalization Process. In: Proceedings of the 9th SIGKDD Conference (2003)
5. Garofalakis, J., Giannakoudi, T., Sakkopoulos, E.: An Integrated Technique for Web Site Usage Semantic Analysis: The ORGAN System. Journal of Web Engineering (JWE), Special Issue Logging Traces of Web Activity 6(3), 261–280 (2007)
6. Haveliwala, T.: Topic-Sensitive PageRank. In: Proceedings of the Eleventh Intl. World Wide Web Conf. (2002)
7. Ma, Z., Pant, G., Sheng, O.: Interest-based personalized search. ACM Trans. Inf. Syst. 25(1) (2007)
8. Makris, C., Panagis, Y., Sakkopoulos, E., Tsakalidis, A.: Category ranking for personalized search. The Data and Knowledge Engineering Journal (DKE), Elsevier Science 60(1), 109–125 (2007)
9. Middleton, S.E., Shadbolt, de Roure, D.C.: Ontological User Profiling in Recommender Systems. ACM Trans. Information Systems 22(1), 54–88 (2004)
10. Miller, G.A.: WordNet: A lexical database for English. Communications of the ACM 38(11), 39–41 (1995)
11. Pedersen, T., Patwardhan, S., Michelizzi, J.: WordNet:Similarity - Measuring the Relatedness of Concepts. In: Proceedings of the Nineteenth National Conference on Artificial Intelligence (AAAI 2004), San Jose, CA, July 2004, pp. 1024–1025 (2004)
12. Wu, Z., Palmer, M.: Verb semantics and lexical selection. In: Proceedings of the 32nd Annual Meeting of the Associations for Computational Linguistics, Las Cruces, New Mexico, pp. 133–138 (1994)
13. Mobasher, B., Cooley, R., Srivastava, J.: Automatic Personalization based on web usage Mining. Communications of the ACM 43(8), 142–151 (2000)
14. MacQueen, J.B.: Some Methods for classification and Analysis of Multivariate Observations. In: Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 281–297. University of California Press, Berkeley (1967)

15. Qiu, F., Cho, J.: Automatic identification of user interest for personalized search. In: Proceedings of the 15th International WorldWide Web Conference, Edinburgh, Scotland, U.K. ACM Press, New York (2006)
16. Tanudjaja, F., Mui, L.: Persona: A contextualized and personalized web search. In: Proc. of the 35th Annual Hawaii International Conference on System Sciences (2002)
17. The Open Directory Project, `http://www.dmoz.org/`
18. The Protégé Ontology Editor and Knowledge Acquisition System, `http://protege.stanford.edu`

# A Comparison of Different Rating Based Collaborative Filtering Algorithms

Fabián P. Lousame and Eduardo Sánchez

Grupo de Sistemas Inteligentes, Dpto. de Electrónica e Computación
Universidade de Santiago de Compostela, 15782 Santiago de Compostela, Spain
fabianpl@usc.es,eduardos@usc.es

**Abstract.** In this paper we present a preliminary work in which different rating based collaborative filtering algorithms are compared in terms of scalability and recommendation quality. Algorithms are tested using a reference database and results show that the selection of one or other algorithm depends on two factors: the scalability of the algorithms and the recommendation quality.

## 1 Introduction

Collaborative filtering techniques are aimed at recommending items by exploiting the preference patterns of a group of users to predict preferences for other items. Unlike other recommendation methods, collaborative filtering tries to predict the utility based on the items previously rated by the users. Several algorithms have been proposed (see Huang et al. [1] and G. Karypis [2], for review) but the general approach to collaborative filtering is based either on generating predictions for user ratings on new items based on a similarity measure among users [3,4], or on computing recommendations by initially exploring item similarities, and then build a recommendation list by aggregating items which are similar to those previously purchased by the user [5,6]. They are the well known user-based and item-based approaches.

Different new approaches and variants have emerged. Wang et al. [7] defined a similarity fusion method to unify user as well as item similarity in a more general similarity measure. Algorithms were also proposed to find users who are similar to the test user by clustering them in segments and treat the recommendation problem as a classification problem [8]. Recent collaborative filtering systems also tried to unify different methods to build more flexible recommender systems. For instance, Huang et al. [9] developed a graph model that provides a generic data representation that supports different recommendation methods. More sophisticated solutions have been proposed to address emerging limitations such data sparsity, scalability and the first-rater problem. For instance, based on hybrid approaches that combine different techniques and information (such user profiles and descriptions of item content) to construct the similarity measures, Liu et al. [10], Shih et al. [11] or Han et al. [12] have proposed new collaborative filtering recommenders.

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 248–256, 2008.

In this preliminary work four different collaborative filtering algorithms are compared and experimental results are presented. The MovieLens dataset [13] was considered to study both the performance and the recommendation quality of the tested algorithms. Similar works have been carried out before (see the comparison presented by Zan Huang et al. [1]) but they mainly differ in the algorithms and datasets. The intention of this work is to compare the algorithms and decide which algorithms are more appropriate, given certain context requirements or limitations.

The paper is organized as follows. In next section we briefly describe the four algorithms we compare. Next, in section 3 we explain the experiments we considered to test and compare the algorithms and in section 4 the results of the experiments are presented and discussed. Section 5 concludes this work.

## 2 Collaborative Filtering Recommendation Algorithms

In a collaborative filtering recommendation problem a user is usually represented as a N dimensional vector of item ratings, where N is the total number of items available. If a test user $k$ has selected[1] item $j$ and has evaluated that item with rating $r_{kj}$, the $j^{th}$ component of the user vector will be $r_{kj}$. The complete rating information is represented in a MxN user-item interaction matrix, $A = UxI$. In the following subsections we briefly describe the approaches to rating based collaborative filtering we have tested.

### 2.1 The User-Based Algorithm

The user-based algorithm generates recommendations by aggregating ratings from similar users. In this approach every user is represented as a vector $\boldsymbol{a_k}$ in which each component $a_{ki}$ denotes the user's rating on item $i$, $r_{ki}$. The algorithm first computes the user similarity matrix and then selects recommendations from the similar users' items. The element $s_{kl}$ in the similarity matrix is obtained by calculating the similarity between users $k$ and $l$ using some vector similarity function. Different similarity functions can be considered but the most common is the cosine measure, extended over all items both users have selected:

$$s_{kl} = \frac{\sum_{i \in B_k \cap B_l} r_{ki} \cdot r_{li}}{\sqrt{\sum_{i \in B_k \cap B_l} r_{ki}^2} \sqrt{\sum_{i \in B_k \cap B_l} r_{li}^2}} \tag{1}$$

Once the similarity matrix is computed, the user-based algorithm selects recommendations from the similar users' items by using various methods. A common method is to rank each item according to how many similar users selected it:

$$\nu_{kj} = \frac{\sum_{l \in U_k} s_{kl} \cdot r_{lj}}{\sum_{l \in U_k} |s_{kl}|} \tag{2}$$

[1] We will use the term 'select' to refer to any action through which the user purchases, reads, etc. an item.

Here $U_k$ denotes the set of users which are similar to user $k$ and $r_{lj}$ the rating of the user $l$ on product $j$.

### 2.2 The Item-Based Algorithm

The item-based algorithm generates recommendations by looking into the set of items the target user $k$ has rated. Given this user and her collection of previously rated items, $B_k$, the item-based algorithm creates a list of recommendations by selecting new items that are similar to this collection. The algorithm first computes how similar a target item $j$ is to the items in the collection $B_k$ and selects the most similar ones. Once the most similar items are found, items are ranked in a recommendation list by using some prediction measure.

The fist step of this algorithm is to compute the similarity between items and select the most similar ones. The item similarity matrix can be obtained using different measures: cosine based similarity, adjusted cosine similarity and correlation based similarity. Among them we will consider the adjusted cosine similarity given by:

$$s_{ij} = \frac{\sum_{k\in U}(r_{ki} - \bar{r_i}) \cdot (r_{kj} - \bar{r_j})}{\sqrt{\sum_{k\in U}(r_{ki} - \bar{r_i})^2}\sqrt{\sum_{k\in U}(r_{kj} - \bar{r_j})^2}} \quad (3)$$

Here $r_{ki}$ represents the rating of user $k$ on item $i$ and $\bar{r_i}$ the average rating of the $i^{th}$ item.

Though other measures are also possible (see [5]), items in the recommendation list are usually ranked with a prediction measure computed by taking a weighted average over all user's ratings for items in the collection $B_k$:

$$\nu_{kj} = \frac{\sum_{i\in B_k} r_{ki} \cdot s_{ij}}{\sum_{i\in B_k} |s_{ij}|} \quad (4)$$

### 2.3 The Item-to-Item Algorithm

The item-to-item algorithm was proposed by Greg Linden et al. [6] to produce recommendations in real time, to scale to massive datasets and to generate high-quality recommendations.

The item-to-item collaborative filtering algorithm matches each of the user's selected and rated items to similar items and then combines those similar items into a recommendation list. The algorithm is quite similar to the item-based proposed by Sarwar et al. [5] but it includes several advantages: (1) the similarity computation is extended only to item pairs with common users (co-ocurrent items) and (2) the recommendation list is computed looking into a small set that aggregates items that were found similar to a certain basket of user selections. These simple modifications make the item-to-item algorithm faster than the item-based.

To determine the most similar match from a given item, the algorithm builds a co-ocurrence matrix by finding items that users tend to select together, $M$.

The similarity between two items $i$ and $j$ is not zero if at least $m+1$ users have selected the pair $(i, j)$, with $m \geq 0$ some predefined threshold. It is possible to compute the similarity between two items satisfying this property in various ways but a common method is to use the cosine similarity described in equation 3. Predictions for new items are computed with equation 4 (see [6] for details).

### 2.4 The Horting Approach

The horting approach, proposed by Charu C. Aggarwal et al. [14], is a graph collaborative filtering algorithm based on the concepts of *horting* and *predictability.* The technique is proposed as a fast, scalable and accurate collaborative filtering algorithm.

The algorithm builds a graph by analyzing two conditions for any two users $k$ and $l$. The first condition is the horting condition and states whether there is enough similarity among each pair of users $(k, l)$ to decide if the behaviour of one user predicts another's or not. It is said that user $k$ horts user $l$ if $card(B_k \cap B_l) \geq min(F \cdot card(B_k), G)$, where $F \leq 1$ and $G$ is some predefined threshold. The second condition is the predictability condition. Two users $k$ and $l$ satisfy this condition if there exists a linear rating transformation $T_{s,t} = s \cdot r + t$ for any pair $(s,\ t)$ of real numbers. The $(s,\ t)$ pair is chosen so that the transformation $T_{s,t}$ keeps at least one value in the rating domain (see [14] for further details on $s$-$t$ value pair restrictions). User $l$ predicts user $k$ if user $k$ horts user $l$ and if there exists a pair $(s,\ t)$ such that the expression 5 is satisfied, with $U$ a positive real number.

$$\frac{\sum_{j \in B_k \cap B_l} |r_{kj} - T_{s,t}(r_{lj})|}{card(B_k \cap B_l)} < U \tag{5}$$

To compute the recommendation list a directed graph is considered. Each arc between users $k$ and $l$ represents that user $l$ predicts user $k$ and therefore it has associated a linear transformation $T_{s_{k,l}, t_{k,l}}$. Using an appropriate graph search algorithm a set of optimal directed paths between user $k$ and any user $l$ that selected item $j$ can be constructed. Each directed path allows a rating prediction computation based on the composition of transformations. For instance, given the directed graph $k \rightarrow l_1 \rightarrow ... \rightarrow l_n$ with predictor values $(s_{k,1}, t_{k,1}), (s_{1,2}, t_{1,2}), ..., (s_{n-1,n}, t_{n-1,n})$ the predicted rating of item $j$ will be $T_{s_{k,1}, t_{k,1}} \circ T_{s_{1,2}, t_{1,2}} \circ ... \circ T_{s_{n-1,n}, t_{n-1,n}}(r_{nj})$. Since different paths may exist, the average of this predicted ratings is computed as the final prediction.

## 3 The Experimental Evaluation

### 3.1 The Test Dataset

We have evaluated the algorithms using a dataset from the MovieLens recommender system [15]. The whole database has about 100.000 preferences that 943 users have specified for 1682 different items. For the scope of this work a selection of 6633 preferences was considered. The sparsity level of this subset was about 96% and the database has ratings in a five point scale, from 1 to 5.

### 3.2 Evaluation Metrics

All the studied algorithms generate a recommendation list in which each item is ranked using a prediction value. For each user we measured the recommendation quality in terms of MAE, a weighted average of the absolute errors $e_i = f_i - y_i$, where $f_i$ is the prediction and $y_i$ the true value. For collaborative filtering, $f_i$ represents the prediction of a user rating and $y_i$ the true user rating. Other measures such precision and recall, the F-measure or the Rank Score [3] are also possible. We have used MAE as our quality evaluation metric since it is the most commonly used and the easiest to interpret.

Besides MAE we considered relevant to measure the performance of the algorithms in terms of time consumed and scalability in time with the number of users. To elaborate the scalability comparison, we split the time consumed by each algorithm in two different contributions: the offline contribution and the online contribution. Table 1 explains each contribution for every algorithm.

**Table 1.** Tasks for each algorithm and execution mode

| Algorithm | Offline opperations | Online opperations |
|---|---|---|
| **user-based** | user similarity computation | prediction computation using a weighed statistical measure |
| **item-based** | item similarity computation | prediction computation using a weighed statistical measure |
| **item-to-item** | item-to-item correlation and item similarity computation | prediction computation using the item-to-item correlation matrix and the item similarity, using a weighed statistical measure |
| **horting** | horting and predictability condition evaluation and computation of prediction parameters $s, t$ for every user | graph building, shortest path search and prediction computation from the directed paths |

### 3.3 Experimental Procedure

To test the recommendation quality of the algorithms, we considered a selection of 2664 preferences specified by 93 users on 664 items. Following the train/test ratio results derived from the work of Sarwar et al. [5] we randomly selected the 80% of the preferences for the training set and the rest for testing purposes. To test the scalability of the algorithms we generated six different datasets, each one with a different number of users, items and preferences. Table 2 summarizes these datasets.

To evaluate the item-to-item algorithm a co-ocurrence threshold $m$ must be specified in order to take advantage of the co-ocurrence matrix. In these experiments we considered $m = 2$. Besides, to compute the prediction for either user-based, item-based or item-to-item algorithms the neighbourhood size was chosen so that the prediction computation is extended over all similar users or

**Table 2.** Datasets used for the scalability test

| Dataset | Preferences | Users | Items |
|---|---|---|---|
| DB1 | 1714 | 60 | 538 |
| DB2 | 2008 | 70 | 584 |
| DB3 | 2286 | 80 | 620 |
| DB4 | 2664 | 93 | 664 |
| DB5 | 4060 | 134 | 792 |
| DB6 | 6633 | 200 | 900 |

items (similarities greater than zero in equations 1 and 3) though better results could be achieved if the neighbourhood size is small.

The horting algorithm requires parameters F, G and U to be fixed. These values were selected to achieve recommendation quality results comparable with the other approaches. They strongly depend on the database properties and their final values were set to $\{F = 0.4, G = 6, U = 0.525\}$.

The prediction step of the horting algorithm also requires a graph search algorithm. Aggarwal et al. [14] considered a shortest path search algorithm but we found more reasonable to use a minimum error path search algorithm. Such algorithm performs a depth search in the graph to find the paths which produce predictions with minimum error. The error can be computed as the accumulative error of all rating transformations $T_{s,t}$ defined by the path that links two users. Paths with an accumulated error greater than a certain fixed threshold were skipped during the search process. This threshold was set to 0.8, which means the 20% of the rating scale.

## 4 Experimental Results

The algorithms were implemented in Java and the experiments were executed in a Intel Core 2 Duo T7200 with a 1GB of DDR2 RAM running Windows XP SP2. The Java Runtime Environment was JDK1.5 and the database engine was MySQL version 5.0.45.

**Recommendation quality.** Figure 1 shows the impact of the different collaborative filtering algorithms in the recommendation quality. It can be observed from these results that both item-based and item-to-item algorithms give better results in this database than user-based or horting. Besides, they have nearly the same MAE, what is consistent with the expected results, since they only differ in the way items are aggregated into a recommendation list.

An advantage of the horting algorithm is that it can be adjusted to produce better recommendations, by decreasing either the error threshold or the predictability threshold U; or by increasing parameters F or G. Despite these changes can produce better predictions, the amount of recommendations per user may decrease and the number of users that do not have predictions will increase.

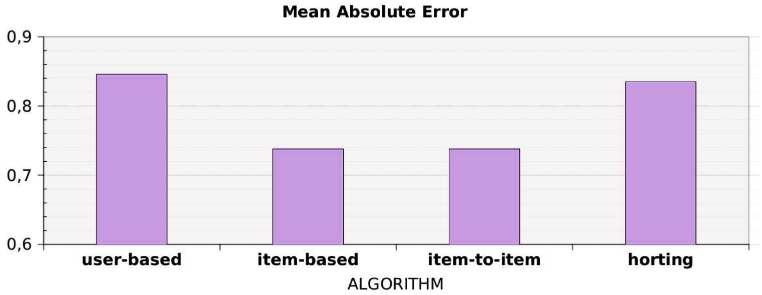


**Fig. 1.** Recommendation quality for MovieLens dataset

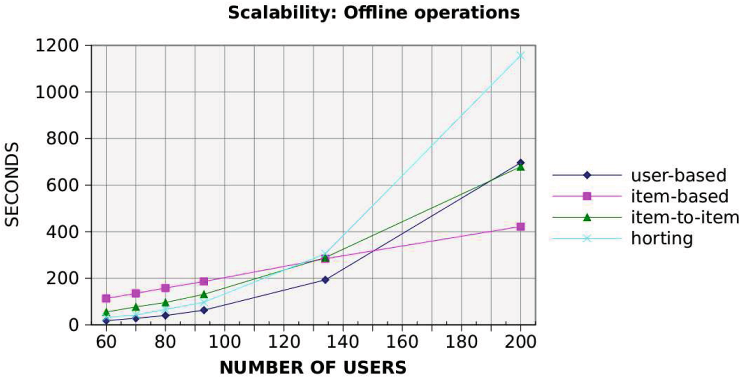


**Fig. 2.** Scalability of the offline operations for the different algorithms

**Offline computation.** Figure 2 plots the scalability results that correspond to the offline operations for the different recommendation algorithms. As we can see the horting algorithm is the less scalable, followed by the user-based. Both the item-based and the item-to-item algorithms seem to scale linearly with the number of users, but the item-to-item algorithm is more time consuming since requires to compute the additional co-ocurence matrix mentioned in section 2.3. Checking horting and predictability conditions is time intensive which makes the algorithm have scalability problems. For 140 users and above, both the user-based and the horting based algorithms have serious scalability problems.

**Online computation.** The online computation scalability results are shown in figure 3. Scalability results are presented in this figure for the item-based, item-to-item and the horting algorithms. The user-based was discarded since computation times are huge compared to the other algorithms. These results show that the algorithm which scales best for the online tasks is the horting based collaborative filtering.

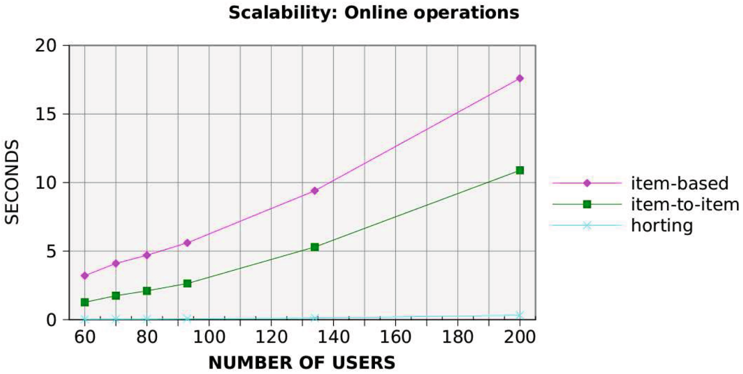


**Fig. 3.** Scalability of the online operations for the most scalable algorithms

## 5 Conclusion

In this paper we have summarized a work in which we evaluated four different algorithms that produce recommendations following a rating collaborative filtering approach. Algorithms were tested using the MovieLens database. The algorithm execution was split into two contributions: the offline and the online computations. From the experiments we can see both the item-based and the item-to-item collaborative filtering algorithms produce good quality recommendations though they may suffer from scalability problems in the online computation step as the number of users/items increases.

At first sight it seems the horting graph approach is the best if we assume the online computation time is the most important factor. Besides, a graph based approach offers the possibility to explore transitive relations among users and build predictions in situations where the interaction matrix is very sparse. The graph search process and predictability functions $T_{s,t}$ can be revised in order to offer better recommendations and make the horting based approach comparable to the item-based or item-to-item algorithms in terms of recommendation quality.

Depending on our requirements we could choose a different algorithm. If the online computation time is the most important factor to decide on a recommendation algorithm, we should choose the horting approach, but if we need recommendation quality, the item-based or the item-to-item algorithms should be considered instead.

## Acknowledgment

The authors would like to thank the support of the ‘Ministerio de Educación y Ciencia (MEC)’ through the grant TSI2007-65677-C02-02.

## References

1. Huang, Z., Zeng, D., Chen, H.: A comparison of collaborative-filtering recommendation algorithms for e-commerce. IEEE Intelligent Systems 22(5), 68–78 (2007)
2. Karypis, G.: Evaluation of item-based top-n recommendation algorithms. In: CIKM 2001: Proceedings of the tenth international conference on Information and knowledge management, pp. 247–254. ACM, New York (2001)
3. Breese, J.S., Heckerman, D., Kadie, C.: Empirical analysis of predictive algorithms for collaborative filtering. In: UAI 1998: Proceedings of the fourteenth conference on uncertainty in artificial intelligence, pp. 43–52 (1998), citeseer.ist.psu.edu/breese98empirical.html
4. Herlocker, J.L., Konstan, J.A., Borchers, A., Riedl, J.: An algorithmic framework for performing collaborative filtering. In: SIGIR 1999: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pp. 230–237. ACM, New York (1999)
5. Sarwar, B., Karypis, G., Konstan, J., Reidl, J.: Item-based collaborative filtering recommendation algorithms. In: WWW 2001: Proceedings of the 10th international conference on World Wide Web, pp. 285–295. ACM, New York (2001)
6. Linden, G., Smith, B., York, J.: Amazon.com recommendations: Item-to-item collaborative filtering. IEEE Internet Computing 7(1), 76–80 (2003)
7. Wang, J., de Vries, A.P., Reinders, M.J.T.: Unifying user-based and item-based collaborative filtering approaches by similarity fusion. In: SIGIR 2006: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 501–508. ACM, New York (2006)
8. Ungar, L., Foster, D.: Clustering methods for collaborative filtering. In: Proceedings of the Workshop on Recommendation Systems, AAAI Press, Menlo Park (1998), citeseer.ist.psu.edu/ungar98clustering.html
9. Huang, Z., Chen, H., Zeng, D.: Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering. ACM Transactions on Information Systems 22(1), 116–142 (2004)
10. Liu, D.-R., Shih, Y.-Y.: Hybrid approaches to product recommendation based on customer lifetime value and purchase preferences. J. Syst. Softw. 77(2), 181–191 (2005)
11. Shih, Y.-Y., Liu, D.-R.: Hybrid recommendation approaches: Collaborative filtering via valuable content information. In: HICSS 2005: Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS 2005) - Track 8, p. 217.2. IEEE Computer Society, Washington (2005)
12. Han, E.-H.S., Karypis, G.: Feature-based recommendation system. In: CIKM 2005: Proceedings of the 14th ACM international conference on Information and knowledge management, pp. 446–452. ACM, New York (2005)
13. Movielens web site (March 2008), http://www.movielens.org/
14. Aggarwal, C.C., Wolf, J.L., Wu, K.-L., Yu, P.S.: Horting hatches an egg: a new graph-theoretic approach to collaborative filtering. In: KDD 1999: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 201–212. ACM, New York (1999)
15. Konstan, J.A., Miller, B.N., Maltz, D., Herlocker, J.L., Gordon, L.R., Riedl, J.: Grouplens: Applying collaborative filtering to usenet news. Communications of the ACM 40(3), 77–87 (1997)

# Application of Agent-Based System for Bioprocess Description and Process Improvement

Ying Gao[1], Katie Kipling[2], Jarka Glassey[2], Mark Willis[2], Gary Montague[2],
Yuhong Zhou[1], and Nigel Titchener-Hooker[1]

[1] Department of Biochemical Engineering, University College London
[2] School of Chemical Engineering and Advanced Materials,
Newcastle University

**Abstract.** An agent-based system framework is developed to provide a flexible environment for analysing bioprocesses based on a whole process understanding. In this system, agent components cooperate with each other in performing their tasks. These include the description of the whole process behaviour, evaluating process operating conditions, monitoring of the operating processes, predicting critical process performance, and providing decision support when coping with process deviations. In all cases the function of the system is to ensure an efficient manufacturing process and to maintain the product quality. The implementation of the agent-based approach is illustrated via a process monitoring scenario.

**Keywords:** Bioprocess modelling, agent-based system, bioprocess interaction, process improvement.

## 1 Introduction

With the pressures to achieve faster development of new biopharmaceutical products and a need to realise cost effective and higher yielding process, there has been increased interest in the use of mathematical models to describe whole bioprocess performance (Zhou, *et. al.,* 1997). Process models can help us to understand the process in greater detail and to allow decisions to be made in a more effective manner. During the process development phase, such models can be used to guide decision-making in selecting the most efficient process sequences and the values of key operating variables. During product manufacture, process deviations are a key concern and tools allowing a swift response will be beneficial in ensuring the quality of products and the efficiency of manufacture are maintained.

Developing bioprocess models is time-consuming and there are several factors that should be considered in order to achieve more efficient process descriptions. Firstly, a bioprocess contains multiple unit operations, and there are strong interactions between the steps which need to be considered in order to investigate the whole process performance (Davies, *et. al.* 2000; Meirels, *et. al.* 2003). Secondly, the time-critical information such as the occurrence of an abnormal situation in the manufacturing process needs to be captured and to be responded to swiftly and effectively. This requires an evaluation of the whole process situation in a timely manner and a

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 257–264, 2008.

capacity to provide constructive decisions to cope with the unexpected situation. Additionally, the iterative process of model development requires a flexible system structure to accommodate the updating of process models and knowledge base.

Traditional bioprocess modelling considers the process units separately, making it difficult to investigate interactions between unit operations, and also limits the capacity to achieve iterative model development. What is required is a system that allows whole process descriptions to be evolved using the available system information and to be updated as new information arises.

Agent-based techniques are ideally suited to such a problem. “An agent is a computer system that is situated in some environment, and that is capable of autonomous action in this environment in order to meet its design objectives” (Wooldridge and Jennings, 1995). A multi-agent system normally consists of a number of agents that work together to solve problems by collaborating and communicating with other software agents in a network (Sycara, 1998). The agents interact with one another, typically by exchanging messages through the computer network infrastructure. These agents share information, knowledge and tasks among themselves. They coordinate and negotiate with each other to achieve common goals. In addition, a multi-agent system can more easily manage the detection and response to important time-critical information that could appear from a number of different information sources (Soler, *et. al.* 2002).

Multi-agent systems provide an excellent environment for modelling whole bioprocesses characterised by multiple unit operations, and with strong interactions between steps. The modular approach of system development enables the user to define a multi-agent architecture providing the ability to integrate unit operation models together. Applying a multi-agent system architecture will also make the system flexible, extendable and reusable, which can support the integration of existing unit operation models, as well as providing for the addition of others when they become available.

The communication ability between agents is well-suited to tackle unit interactions, and the ability to respond to time-critical information allows agents to detect changes in process operations and to represent the dynamic bioprocess operation behaviour. They provide a capacity to monitor time-critical information in the manufacturing process, particularly in detecting deviations in upstream unit operation and predicting the consequences on the subsequent downstream processing whilst providing guidance in terms of possible corrective actions so as to improve process efficiency and ensure product quality.

In this paper, an agent-based system framework is proposed to provide a flexible environment for the necessary integration of bioprocess models to form the whole bioprocess descriptions and to evaluate strategies for plant-wide manufacturing improvements. The architecture of the multi-agent system is introduced in section 2, and the application of the proposed system is illustrated in section 3.

## 2 Framework for Agent-Based Plant-Wide Bioprocess Description

### 2.1 Multi-agent System Architecture

Figure 1 illustrates the multi-agent system architecture proposed in this work. The system comprises a process knowledge base, unit operation models, a group of functional agents and a user interface. The process knowledge base itself contains data and information from the different unit operations along the process sequence. The data sets include experimental data which is obtained from small-scale experiments, and historical manufacturing data. Information on the unit operation conditions and equipment details are also provided in the knowledge base. Data and information can be classified and organised at different levels in the knowledge base for convenient utilization.

The unit operation models can be developed based on a first principles understanding and may use parameter values derived from small scale experiments. Models can also be obtained based on the data-based methods, in which case they will essentially be black-box by nature. Experimental data combined with the first principle models will mimic steady-state unit operation behaviour, and provide for the fast generation of unit operation descriptions. Data-driven models can be developed by applying data mining techniques to the manufacturing data and will result in a dynamic model which can describe and predict manufacturing process behaviour. The combination of these two types of models will lead to a hybrid model which integrates the best features of these two different types of models with improved predictive performance of the manufacturing process and more comprehensive process description ability.

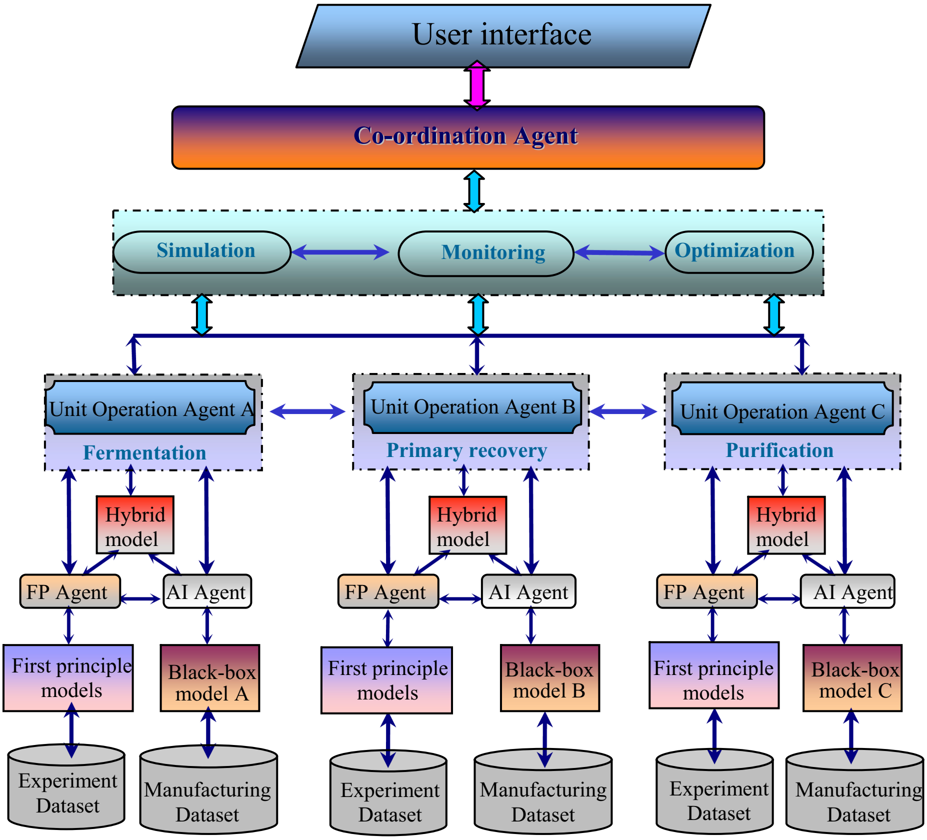


**Fig. 1.** Multi-agent system framework

Agents are organised in a hierarchical architecture within the framework. The lower level agents are able to access the process knowledge base and datasets, and to represent the unit operation behaviour by utilising unit operation models. The higher level agents will be able to integrate unit operations, then communicate and co-operate to simulate the interaction between units and therefore provide a whole process description. The collaboration of agents at different levels would enable the identification of the optimal process operation conditions in order to achieve best overall plant performance, and provide decision-making when coping with various scenarios that may occur in the manufacturing process.

The hierarchical architecture of the agent-based system has the benefit of making the agents work most efficiently. It is easy to add new agents into the system, and the modification can be done without significant changes in the system structure. Updating and adding new information and unit operation models to such a system can also be achieved easily. This facilitates the timely assembling of process models and the updating of process descriptions, especially when the process information is updated or new process models become available to the knowledge base, i.e. as process data accumulates with the running of more manufacturing campaigns.

## 2.2 Function of Agents

The proposed agent-based system comprises a coordination agent, unit operation agents and a group of first principle (FP) agents and artificial intelligent (AI) agents.

The *Co-ordination agent* is responsible for task decomposition and coordinating the activities of the sublevel agents. It receives the tasks specified by the user, e.g. whole process description, process monitoring, process performance prediction, process optimization, *etc.*. It decomposes the tasks and assigns the tasks to the sub-level agents to execute. Co-ordination agent also guide the decision-making based on the decision rules in coping with process variations, and finally notify the result to the user.

*Unit operation agents* represent the individual unit operation behaviour in the bioprocess, such as fermentation, centrifugation and chromatography. They work as a team in the evaluation of process conditions (i.e. the fermentation harvest time, centrifugation flowrate and the loading conditions of chromatography) for improving the whole process performance. Unit operation agents can carry out different activities according to the requirement of the Co-ordination agents, which include process description based on process models, searching for feasible process operation conditions, detecting process deviations against historical process operation data, *etc.*. Unit operation agents can also choose specific unit operation models to describe the process operation performance, e.g. first principle models, data-based models, hybrid models, based on the requirement of the Co-ordination agent and the availability of the models.

*FP agents* access both the experimental dataset and the first principle models developed using experimental data. *AI agents* are designed to perform the tasks of process simulation and process performance prediction using black-box models developed based on the manufacturing data.

FP agents and AI agents act on the information received from the Unit operation agents. If a hybrid model is required for a more sophisticated process description, then both the FP and AI agents will be utilised to derive a hybrid model, under the supervision of the Unit operation agents.

## 3 Application of the Agent-Based System Framework

The proposed agent-based system framework can be applied during process development or once manufacturing has commenced. In the process development stage, the proposed agent–based system can be used to evaluate the design space for process operations, and to identify the optimal process operation performance. In the manufacturing stage, the agent-based system can be applied during process monitoring to identify abnormal process operation events and provide suggestions to cope with the deviations. In this section, we discuss the application of the agent-based system in a process monitoring scenario. A typical intracellular protein production process is used as an example to illustrate the system implementation.

### 3.1 Application Scenario

Figure 2 shows part of a process flowsheet for the production and isolation of an intracellular product alcohol dehydrogenase (ADH) from a *S. cerevisiae* strain. In this process, the micro-organism is cultured in a fed-batch fermenter. Cells are harvest using a high-speed disc-stack centrifuge, and then re-suspended and disrupted in a homogeniser to release the intracellular protein. The same kind of centrifuge is used to remove the cell debris. Subsequent processing of the protein supernatant will include multiple chromatographic purification and ultrafiltration steps for product concentration.

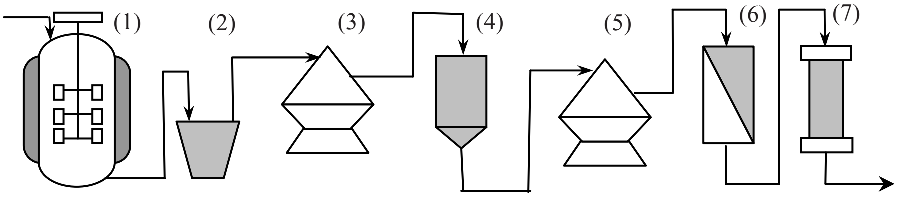


**Fig. 2.** Flowsheet of an intracellular product recovery process. (1) Fermenter, (2) Harvest Tank, (3) Disc-stack centrifuge, (4) Homogeniser, (5) Disc-stack centrifuge, (6) Microfiltration, (7) Chromatography column

In this process, upstream unit operations have a pronounced impact upon the effectiveness of downstream recovery. For example, substrate concentration, feeding strategy and harvest time are important factors for the cell growth, since they affect the biomass concentration, protein level, the size of cells and the ability of the cells to be disrupted during the homogenisation step. In the homogenisation unit, the operation pressure and number of passages can create variability in the centrifuge feed, thus affecting the following centrifugation performance. To obtain the target clarification level, the centrifuge flowrate should be adjusted to account for the number of passages and pressure in the homogenisation step. If the whole process productivity needs to be considered as a target for the manufacturing process, the trade-off between all of the unit operations has to be investigated.

In the manufacturing scenario, process operation conditions and the key process performance parameters, such as cell growth profile in the fermentation process, solid concentration in the centrifugation supernatant, etc. should be controlled within certain boundaries in order to maintain product quality and process consistency. The proposed agent-based system can be used to monitor the process operation performance, and to set an alert if any deviation is detected during the manufacturing process. The system can also help to predict trends in the subsequent process units under alterative operation conditions based on the process models. With the application of decision-rules and optimisation strategies, agents in the system will also collaborate to provide strategies to maintain the efficiency of the manufacturing process. This is driven by agent-based decision-making and the proposal of alternative unit operation conditions to cope with any deviations detected.

### 3.2 System Implementation

The proposed agent-based framework was implemented using JADE as the agent development platform. FIPA ACL was used as the communication language in JADE. An ontology and domain specific vocabulary was defined in the multi-agent system to facilitate the communication between agents and interpret the content information of the messages exchanged between agents. With the communication language and the ontology, agents will be able to understand each other and to communicate effectively to exchange and share information, and achieve timely decisions in dealing with various scenarios.

The ADH process is used as an example to illustrate the implementation of the agent-based system. Unit operation models for fermentation, centrifugation, homogenization and chromatography operations were developed in Matlab. An agent system was set up in which a group of Unit operation agents were used to represent the unit operations along the process sequence. Additionally a Coordination agent was used to coordinate the unit operations, and a User agent to communicate between the user and the agent system, to get user commands and present the results to the user.

Here we take the monitoring of fermentation process as an example. The main kind of deviation that may be expected in the fermentation would be whether the culture growth rate changes relative to historical trends. This affects subsequent processing of the batch. Hence if the final biomass concentration deviates from the target level, may necessitate a change of the downstream unit operation conditions. In a manufacturing scenario, set points will exist for each stage of the process and it is the function of the agents to cooperate in proposing appropriate control actions to bring the problem back within the defined process envelop.

A 1000L yeast fermentation was studied in this example. Set points for unit operation conditions and the key control parameters in this process are list in Table1.

**Table 1.** Set points for unit operations and key control parameters

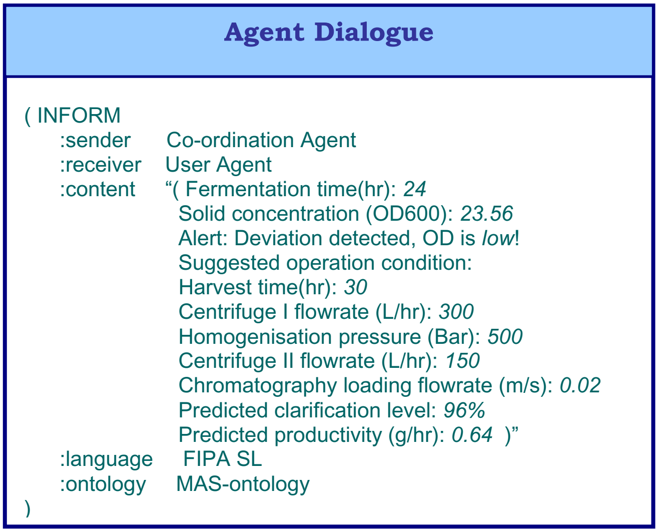
| Set points of unit operation conditions | Key control parameters |
|---|---|
| Fermentation time(hr): 30 | Biomass concentration of fermentation broth: > 120g/L |
| Centrifugation I Flowrate (L/hr):300 | |
| Homogenization Pressure (Bar): 400 | Supernatant clarification of centrifuge II: >95% |
| Centrifugation II Flowrate (L/hr): 200 | |
| Chromatography loading flowrate (m/s):0.02 | Productivity: >0.5 gADH/hr |

As the fermentation progresses, data such as the on-line OD600 which can be used to infer the level of biomass concentration was evaluated using the multi-agent system.

Agents in the multi-agent system carried out the following activities in executing the process monitoring task:

- The Fermentation agent compares the current operation data with historic manufacturing data based on request from the Co-ordination agent.
- If the deviation is within the normal operation boundary (less than 10%), the Fermentation agent will report that operation is satisfactory. If more than a 10% derivation is detected, the Fermentation agent will trigger an alert. In such an event, the Fermentation agent collaborates with the other Unit operation agents to predict the downstream unit operation behaviour at the set point of each unit operation condition.
- The Coordination agent evaluates the whole process operation performance to check whether the desired level of process productivity will be achieved. If by operating the process at the identified set points will not satisfy this requirement, the coordination agent will suggest changes to the unit operation conditions according to a set of decision rules.
- The Co-ordination agent will eventually propose possible actions to the user in order to accommodate the deviation.

In this scenario, an alert is triggered by the Fermentation agent at 24 hr of fermentation. A greater than 10% deviation of OD600 is detected from the set point. Based on the process model prediction and decision rules, the agent system suggests alternative



**Agent Dialogue**

```
( INFORM
    :sender      Co-ordination Agent
    :receiver    User Agent
    :content     "( Fermentation time(hr): 24
                   Solid concentration (OD600): 23.56
                   Alert: Deviation detected, OD is low!
                   Suggested operation condition:
                   Harvest time(hr): 30
                   Centrifuge I flowrate (L/hr): 300
                   Homogenisation pressure (Bar): 500
                   Centrifuge II flowrate (L/hr): 150
                   Chromatography loading flowrate (m/s): 0.02
                   Predicted clarification level: 96%
                   Predicted productivity (g/hr): 0.64  )"
    :language     FIPA SL
    :ontology    MAS-ontology
)
```

**Fig. 3.** Message between Co-ordination and User agent in the event of a deviation being detected. The agent proposes a possible choice of action to follow in this event.

unit operation conditions to cope with the lower biomass concentration detected in the fermentation process. Higher homogenisation pressure is suggested to improve the protein release, and simultaneously altering the centrifuge II flowrate to maintain the clarification level. A typical dialogue between the Co-ordination and the User agents is shown in Figure 3, which illustrates a summary of the process monitoring result, and the suggested corrective actions to cope with the deviation detected in the fermentation process.

## 4 Conclusion

A systematic framework for the generation of an agent-based whole bioprocess description is proposed in this paper. The agent-based framework provides a flexible environment for the necessary integration of process models for the description of whole process behaviour and the inclusion of the interactions between unit operations. In such a system, agent components run on top of process models and datasets, they communicate with each other to exchange process information, cooperate in performing the tasks for whole process description, evaluate unit operation conditions, and identify process optimization strategies. The system also has the capacity to monitor the manufacturing process, to predict process performance in the case of process variations, and assist in decision–making to improve process efficiency within the process design space.

The application of the agent system approach is discussed via a selected scenario which demonstrates how such a framework can provide for the better integration of process operations for the purposes of plant-wide process description and process improvement. The implementation of the multi-agent system is illustrated based on a typical intracellular protein production process.

## References

1. Davies, J.L., Baganz, F., Ison, A.P., Lye, G.J.: Studies on the interaction of fermentation and microfiltration operations: Erythromycin recovery from Saccharopolyspora erythraea fermentation broths. Biotechnology and Bioengineering 69(4), 429–439 (2000)
2. Meirels, M., Lavoute, E., Bacchin, P.: Filtration of a bacterial fermentation broth: harvest conditions effects on cake hydraulic resistance. Bioprocess and Biosystems Engineering 25(5), 309–314 (2003)
3. Soler, J., Julian, V., Rebollo, M., Carrascosa, C., Botti, V.: Towards a real-time MAS architecture. In: Proceedings of Challenges in Open Agent Systems, AAMAS 2002, Bolonia, Italia (2002)
4. Sycara, K.P.: Multiagent Systems. Artificial Intelligence Magazine 19(2), 79–92 (1998)
5. Wooldridge, M., Jennings, N.R.: Intelligent agents: Theory and practice. The Knowledge Engineering Review 10(2), 115–152 (1995)
6. Zhou, Y.H., Holwill, I.L.J., Titchener-Hooker, N.J.: A study of use of simulation for the design of integrated downstream processes. Bioprocess Engineering 16, 367–374 (1997)

# A Dynamic Bayesian Network to Represent a Ripening Process of a Soft Mould Cheese

Cédric Baudrit[1], Pierre-Henri Wuillemin[2], Mariette Sicard[1], and Nathalie Perrot[1]

[1] UMR782 Génie et Microbiologie des Procédés Alimentaires. AgroParisTech, INRA, F-78850 Thiverval-Grignon
[2] Laboratoire d'Informatique de Paris VI (CNRS UMR7606) F-75016 Paris, France
{cbaudrit,msicard,nperrot}@grignon.inra.fr,pierre-henri.wuillemin@lip6.fr

**Abstract.** Available knowledge to describe food processes has been capitalized from different sources, is expressed under different forms and at different scales. To reconstruct the puzzle of knowledge by taking into account uncertainty, we need to combine, integrate different kinds of knowledge. Mathematical concepts such that expert systems, neural networks or mechanistic models reach operating limits. In all cases, we are faced with the limits of available data, mathematical formalism and the limits of human reasoning. Dynamical Bayesian Networks (DBNs) are practical probabilistic graphic models to represent dynamical complex systems tainted with uncertainty. This paper presents a simplified dynamic bayesian networks which allows to represent the dynamics of microorganisms in the ripening of a soft mould cheese (Camembert type) by means of an integrative sensory indicator. The aim is the understanding and modeling of the whole network of interacting entities taking place between the different levels of the process.

**Keywords:** Knowledge integration, food processing, cheese ripening, Dynamic Bayesian Networks, uncertainty.

## 1 Introduction

Available knowledge to describe microbiological and physicochemical changes present during food processes has been capitalized from different sources (experts, literature, sensors). It is expressed under different forms (database, expert opinions, mechanistic models ...) and at different scales (microbial view, sensory view). The reconstruction of dynamics and interactions between the different levels of knowledge is a complex problem due to the heterogeneity of knowledge.

A part of our work is to propose mathematical tools which allow to integrate this capitalized heterogeneous knowledge in order to describe and represent the principal kinetics of observed phenomena by taking into account uncertainty relative to knowledge process (see figure 1.a). The reconstruction of the puzzle of knowledge could lead to a better understanding of the whole network of interacting entities taking place between different levels of food processes.

Mathematical concepts based on expert systems [11, 12] or "‘black box"’ models [13, 18] such that neural networks or "‘white box"’ such that mechanistic models [2, 20]

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 265–272, 2008.

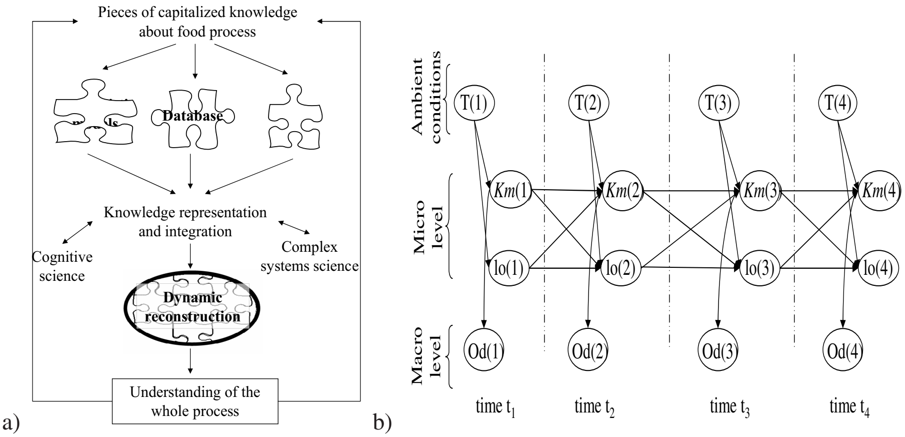


**Fig. 1.** a) Knowledge integration to represent the dynamics of food processes. b) A Simplified representation of the coupled dynamic of *K. marxianus* (*Km*) growth with lactose consumption (lo) influenced by temperature (T) and involving odor changes (Od) for four time slices by means of a DBN.

either reach their limits of feasibility or are not suited. The first one requires expert knowledge sometimes too important for the capacity of human reasoning. The second one is not interpretable and is unable to explain any results that it obtains (*i.e.* the rules of operation are completely unknown). The last one provides fragmented knowledge and does not often allow to take into account, for instance, sensory properties. In the three cases, we are faced with the limits of interpretability, mathematical formalism and the limits of human reasoning. This leads to combine, integrate different kinds of knowledge coming from different sources collected at different scale [1].

With this intention, Dynamical Bayesian Networks (DBNs) [17] are practical probabilistic graphic models to represent dynamical complex systems tainted with uncertainty. The graphical structure of the network provides an intuitively appealing interface by which humans can model highly-interacting sets between variables and provides a qualitative representation of knowledge. The use of probabilities enables to take into account uncertainty pertaining to the system by quantifying dependence between variables in the form of conditional probabilities. The main interest to use DBNs is to permit to combine expert knowledge with experiments through data acquisitions at different levels and scales of knowledge. In addition to the possibility to represent dynamic processes, the use of DBNs allows to elaborate new models by modifying the structure of network and to combine different networks together in order to obtain a new model.

The cheese, during ripening, represents an ecosystem and a bio-reactor difficult to apprehend in the whole. Cheese ripening remains a complicated process to control where operator's evaluation and reasoning have a decisive role. Micro-organisms are the major actors in the process of cheese ripening [7, 8]. The degradations of substrate components can be carried out simultaneously or successively and involve a change of the rheological characteristics of cheese. The aim of this paper is to present a simplified model based on Dynamic Bayesian Networks which allows to represent the dynamics of microorganisms in the ripening of a soft mould cheese (Camembert type) by means of an integrative sensory indicator.

## 2 Basic Notions of Dynamic Bayesian Networks (DBNs) [17]

DBNs are classical Bayesian networks [19] in which nodes $\{X_i(t), i = 1 \dots N\}$, representing random variables, are indexed by time $t$. They provide a compact representation of the joint probability distribution for a finite time interval $[1, \tau]$ defined as follows:

$$P(X(1), \dots, X(\tau)) = \prod_{i=1}^{N} \prod_{t=1}^{\tau} P(X_i(t)|\mathrm{Pa}(X_i(t))) \tag{1}$$

where $X(t) = \{X_1(t), \dots, X_N(t)\}$, $\mathrm{Pa}(X_i(t))$ denotes the parents of node $X_i(t)$ and $P(X_i(t)|$ $\mathrm{Pa}(X_i(t)))$ denotes the conditional probability function associated with the random variable $X_i(t)$ knowing $\mathrm{Pa}(X_i(t))$. This probability represents the beliefs about possible trajectories of the dynamic process $X(t)$. Figure 1.b illustrates a DBN representing a coupled dynamic of a micro organism (*Kluyveromyces marxianus*, noted *Km*) with its substrate consumption (lactose noted lo) influenced by the temperature (T) and causing odour evolutions (Od) during the ripening of Camembert-type cheese. In the structure elaboration of DBNs by means of interactions with the experts of food science, temporal arcs can be seen as the persistence of a phenomenon during time whereas others arcs can be seen as causal influences between variables. Each static network is called a time slice of the DBN. In our framework, we assume the first-order homogeneous Markov property which means that the parents of a variable in time slice t must occur in either slice t or t-1 and that conditional probabilities are time-invariant. DBNs can be then specified simply by giving two slices and the link between them.

DBNs provide very useful tools to combine expert knowledge with data at different levels and scales of knowledge. Indeed the structure of the model can be built explicitly on the basis of the expert knowledge and the parameters of the model (conditional probability functions) can be learnt automatically without a priori knowledge on the basis of a set of data (called parameters learning).

### 2.1 Parameters Learning

The techniques for learning DBNs are mostly extensions of the techniques for learning BNs. They can relate to either the structures of graph (*i.e.* topology) or the parameters of DBNs (*i.e.* conditional probability distributions) or both joined together. According to the figure case, different methods exist to learn the structure or the parameters from substantial and/or incomplete data. Learning DBNs is a huge research subject, we invite readers to read the following papers [4,5] for further details. According to the concern to interact with experts, the topology of the graph will be obtained from expert opinions. Let $\theta^t_{ijk}$ be the probability that $X_i(t) = x_k$ given that its parents have instantiation $y_j$, *i.e.* $\theta^t_{ijk} = P(X_i(t) = x_k|Pa(X_i(t) = y_j)$. The most used and simplest method, which will be used in this paper, is to estimate $\theta^t_{ijk}$ by the occurrence rate of the event $(X_i(t) = x_k, Pa(X_i(t)) = y_j)$ in the training data $S = \{S_1, \ldots, S_Q\}$ which contains $Q$ sequences:

$$\theta^t_{ijk} = \frac{N^t_{ijk}}{\sum_l N^t_{ijl}} \tag{2}$$

where $N^t_{ijk} = \sum_{m=1}^{Q} I(X_i(t) = x_k, Pa(X_i(t)) = y_j|S_m)$ and $I(X_i(t) = x_k, Pa(X_i(t)) = y_j|S_m) = 1$ if the event $(X_i(t) = x_k, Pa(X_i(t)) = y_j)$ occurs in case $S_m$.

### 2.2 Inference (Knowledge Propagation)

Exploitation of such DBNs consists in "query" expressed as conditional probability. The most common task we wish to solve is to estimate the marginal probabilities $P(X_i(t)|\{X_i(t) = x_i(t), (i,t) \in [\![1,N]\!] \times [\![1,\tau]\!]\})$ where $\tau$ is the length of observable data $x_i(t)$. Inference consists in computing the probability of each state of a variable when we know the state taken by other variables. In general, DBN inference is performed using recursive operators that update the belief state of the DBN as new observations become available [17].

- If $\tau = t$, inference is filtering where the goal is to recursively estimate the belief state using Bayes' rule.
- If $\tau > t$ this is smoothing where we look for to estimate the state of the past, given all the evidence up to the current time.
- If $\tau < t$ this is prediction where we might want to predict the future.

## 3 Cheese Ripening Description

For soft-mould cheese the most important biochemical phenomena occur during ripening. There exists relationships between microbiological and physicochemical changes which depend on environmental conditions (*e.g.* temperature, relative humidity ...) [14] and influence the quality of ripened cheeses [10, 15]. The ripening expert is able to explain a part of the complex reactions that are taking place in the real cheese through his perception of the quality changes. Based on this qualitative understanding of the process he generally makes his control decisions. Model cheeses were prepared from pasteurized milk inoculated with *Kluyveromyces marxianus* (*Km*), *Geotrichum candidum* (*Gc*), *Penicillium camemberti* (*Pc*) and *Brevibacterium auriantiacum* (*Ba*) under aseptic conditions (detailed in [15]).

### 3.1 Step of Ripening: Sensory Description

Experts use their senses to follow cheese ripening and they probably aggregate (which so called "'chunk"' in cognitive science [6]) these informations to regulate the evolution of the process. We look for information aggregation as much as possible. During interview with expert-operators, we observed that sensory information taken at line were chunked to represent in four phases the cheese ripening evolution:

- **Phase 1** is defined by the surface humidity evolution of cheese corresponding to drying process. At the beginning, the surface of cheese presents a very wet aspect and diminishes until a nearly dry aspect characterizing the first phase of ripening.
- **Phase 2** is defined by the apparition and the evolution of *P. camemberti*-coat (*i.e* the white-coat at the surface of cheese), the first change of color and the "mushroom" odor development.
- **Phase 3** is defined by the increase of the thickening of the creamy under-rind. *P. camemberti* cover all the surface of cheeses and the color is light brown.
- **Phase 4** is defined by strong ammoniac odor perception and the dark brown aspect of the rind of cheese.

### 3.2 Microbial Behavior during Ripening

*K. marxianus* is one of the dominant species in the yeast flora of Camembert cheeses and has a key role in ripening. One of its principal activity is the fermentation of lactose (noted lo) [7, 8]. Three dynamics are apparent in the timeline of *K. marxianus* growth [15, 14]. Firstly, there is an exponential growth during about five days what corresponds to a decrease of lactose concentration. Secondly, the concentration of *K. marxianus* remains constant during about fifteen days and thirdly decreases slowly. *G. candidum* is used as a starter in the dairy industry and plays a key role ripening because it contributes to the development of flavor, taste and aroma of cheeses [3, 16]. One of its principal activities is the consumption of lactate (noted la). Three dynamics are apparent in the timeline of *G. candidum* growth [15, 14]. Firstly, there is a latency period during about three days. Secondly, there is an exponential growth what corresponds to a decrease of lactate concentration and thus an increase of pH. Thirdly, the concentration of *G. candidum* remains constant to the end of ripening. There is a lack of well suited measure to follow the growth of *P. camemberti*. For instant, the sampling method destroy the mycelia of *P. camemberti*. So the viable cell counted are not directly correlated with the growth of *P. camemberti* on the surface of cheese. We thus decide that this micro-organism will not be taken into account in our modelization. *B. auriantiacum* is an sensitive bacteria growing slowly after a latence period about one week and the population becomes more stable at the end of ripening.

## 4 First Results

From the knowledge summarized above about ripening process, we propose to build a dynamic bayesian network (see Fig. 2) representing the dynamic behaviors of each variable, not from a microscopic but macroscopic point of view: each dynamic is only characterized by the phase in which the network is at time t. From the analysis of experiments and according to expert opinions, we identified four relevant variables (the derivative of pH, la, *Km* and *Ba*) allowing to predict phase at time $t + 1$. After parameters learning (see Section 2.1) obtained from trials run at different relative humidities varying between 88% and 98% and temperatures varying from $8^o$C to $16^o$C, we compare the results of the model with one trial carried out at 88%, $16^o$C non available in learning database. The Figure 3.a allows to show the probability

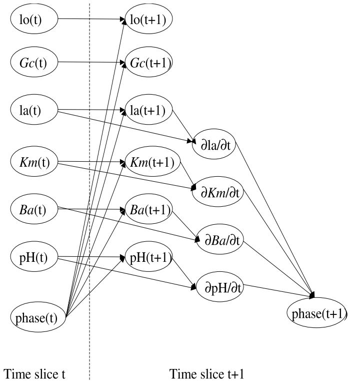


**Fig. 2.** DBN representing the dynamic of variables depending on the observation of ripening phases

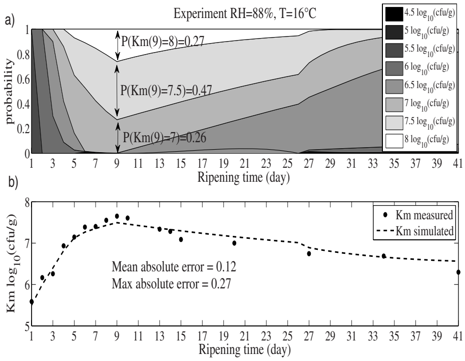


**Fig. 3.** a) Probability of *Km*(t) by observing the phase at each time step. b) Mean of *Km*(t) versus *Km*(t) measured.

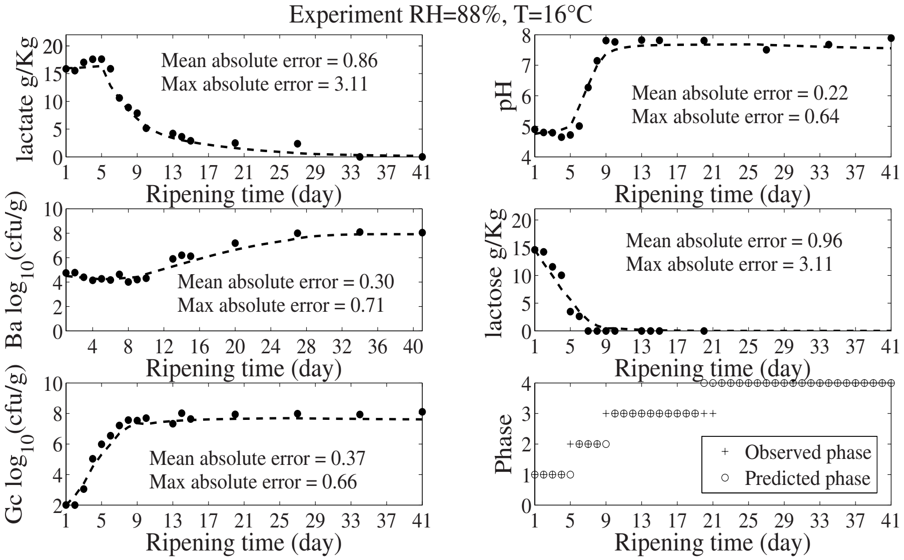


**Fig. 4.** Probability of la(t), pH(t), *Ba*(t), lo(t), *Gc*(t) and predict phase by observing the phase at each time step during ripening with a relative humidity of 88% and a temperature temperature $16^o$C

$P(Km(\tau + \delta t)|Km(1), Gc(1), Ba(1), \mathrm{lo}(1), \mathrm{la}(1), \{\mathrm{phase}(t), t \in [\![1, \tau]\!]\}), \forall \tau \in [\![1, 40]\!]$ and $\delta t = 1$ day, obtained from the DBN where $(Km(1), Gc(1), Ba(1), \mathrm{lo}(1), \mathrm{la}(1), \{\mathrm{phase}(t), t \in [\![1, \tau]\!]\})$ are observed evidences coming from the ripening carried out at 88% and $16^o$C. The Figure 3.a shows, for instance, that the probability for that *Km* is equal to $3.16 \times 10^7$ **c**olony **f**orming **u**nit per **g**ram of wet cheese (*i.e.* about 7.5 in decimal logarithm scale noted "$\log_{10}$(cfu/g)") at time of 9 days is equal to 47% (*i.e.* P(*Km*(9)=7.5)=47%). From the Figure 3.a, we estimate the mean of *Km*(t) (see dotted line in Figure 3.b) and we test the adequation of our model with data coming from trial and thus its predictive charactere. For example, the simulated kinetic of *Km* presents a mean error of 0.12 $\log_{10}$(cfu/g of wet cheese) with the measured kinetic which is lower than measure error admitted by expert: 0.5 $\log_{10}$(cfu/g of wet chesse). In the same way, the Figure 4 shows the simulated mean evolution of la, pH, *Ba*, lo, *Gc* (in dotted line) during ripening at a relative humidity of 88% and a temperature of 16$^o$C. They allow to compare results obtained from simulation with raw data and highlight the predictive character of model in particular about the prediction of the ripening phase from derivatives. We can see that the model correctly predicts the phase in 90% of cases (see predicted phases ∘ in Figure 4).

## 5 Conclusion and Discussions

According expert knowledge, we defined a first dynamic bayesian network allowing to describe a network of interactions taking place between variables at different scales during the ripening of a soft mould cheese. The first results show that the model describes and predicts the kinetics of the characteristic variables of cheese ripening according to the observation of ripening phases. Moreover, it allows to correctly predict the phase at time $t+1$ knowing the phase at time $t$. DBNs thus presents interesting and promising results to take explicitly into account the fragmented and heterogeneous knowledge on the dynamics of the process. However, further tests must be to carry out at different process controls in order to test the robustness of model and to validate it. In further studies, we hope to complex the model by integrating: (1) specific sensory indicators characterizing ripening phases, (2) more precise dependencies between microscopic variables and the knowledge of proteolysis and lipolysis activities. However, according to the complexity of microbiological and/or physicochemical activities in food processes, available knowledge is often tainted with vagueness, imprecision and incompleteness. In order to integrate this kind of knowledge, we can imagine to have to combine RBDs with recent theories like possibility theory [9], Dempster-Shafer theory [21] allowing to process imprecision and incompleteness.

## Acknowledgement

We thank Leclercq-Perlat, M.N., to have provided the data and Picque, D., Corrieu, G., and Spinnler, E. for their expertise.

## References

1. Allais, I., Edoura-Gaena, R.B., Gros, J.B., Trystram, G.: How human expertise at industrial scale and experiments can be combined to improve food process knowledge and control. Food Res. Int. 40, 585–602 (2007)
2. Aldarf, M., Fourcade, F., Amrane, A., Prigent, Y.: Substrate and metabolite diffusion within model medium for soft cheese in relation to growth of Penicillium camembertii. J. Ind. Microbiol. Biotechnol. 33, 685–692 (2006)

3. Boutrou, R., Guéguen, M.: Interests in Geotrichum candidum for cheese tencnology. Int. J. Food Microbiol. 102, 1–20 (2005)
4. Buntine, W.L.: Operations for learning with graphical models. J. AI Res., 159–225 (1994)
5. Buntine, W.: A guide to the literature on learning probabilistic networks from data. IEEE Trans. On Knowledge And Data Eng. 8, 195–210 (1996)
6. Chase, W.G., Simon, H.A.: Perception in chess. Cognitive Psychologie 4, 55–81 (1973)
7. Choisy, C., Desmazeaud, M.J., Gripon, J.C., Lamberet, G., Lenoir, J.: La biochimie de l'affinage. In: Eck, A., Gillis, J.C. (eds.) Le fromage, pp. 86–105. Tec Doc Lavoisier, Paris (1997)
8. Choisy, C., Desmazeaud, M.J., Gueguen, M., Lenoir, J., Schmidt, J.L., Tourneur, C.: Les phénomènes microbiens. In: Eck, A., Gillis, J.C. (eds.) Le fromage, pp. 377–446. Tec Doc Lavoisier, Paris, France (1997)
9. Dubois, D., Nguyen, H.T., Prade, H.: Possibility theory, probability and fuzzy sets: misunderstandings, bridges and gaps. In: Dubois, D., Prade, H. (eds.) Fundamentals of Fuzzy Sets, pp. 343–438. Kluwer, Boston (2000)
10. Gripon, A.: Mould-ripened cheeses. In: Fox, P.F. (ed.) Cheese: Chemistry, Physics and Microbiology, vol. 2, pp. 111–136. Chapman & Hall, London (1993)
11. Ioannou, I., Perrot, N., Curt, C., Mauris, G., Trystram, G.: Development of a control system using the fuzzy set theory applied to a browning process - a fuzzy symbolic approach for the measurement of product browning: development of a diagnosis model - part I. Journal Of Food Engineering 64, 497–506 (2004)
12. Ioannou, I., Perrot, N., Mauris, G., Trystram, G.: Development of a control system using the fuzzy set theory applied to a browning process - towards a control system of the browning process combining a diagnosis model and a decision model - part II. J.Food Eng. (64), 507–514 (2004)
13. Jimenez-Marquez, S.A., Thibault, J., Lacroix, C.: Prediction of moisture in cheese of commercial production using neural networks. Int. Dairy J. 15, 1156–1174 (2005)
14. Leclercq-Perlat, M.N., Picque, D., Riahi, H., Corrieu, G.: Microbiological and Biochemical Aspects of Camembert-type Cheeses Depend on Atmospheric Composition in the Ripening Chamber. J. Dairy Sci. (89), 3260–3273 (2006)
15. Leclercq-Perlat, M.N., Buono, F., Lambert, D., Latrille, E., Spinnler, H.E., Corrieu, G.: Controlled production of Camembert-type cheeses. Part I: Microbiological and physicochemical evolutions. J. Dairy Res. (71), 346–354 (2004)
16. Lenoir, J.: The surface flora and its role in the ripening of cheese. Int. Dairy Fed. Bull. 171, 3–20 (1984)
17. Murphy, K.P.: Dynamic Bayesian Networks: Representation, Inference and learning. Ph.D. thesis, University of California, Berkeley (2002)
18. Ni, H.X., Gunasekaran, S.: Food quality prediction with neural networks. Food Technology 52, 60–65 (1998)
19. Pearl, J.: Probabilistic Reasoning in Intelligent systems: Networks of Plausible Inference. Morgan Kaufmann, San Diego (1988)
20. Riahi, M.H., Trelea, I.C., Leclercq-Perlat, M.N., Picque, D., Corrieu, G.: Model for changes in weight and dry matter during the ripening of a smear soft cheese under controlled temperature and relative humidity. International Dairy Journal 17, 946–953 (2007)
21. Shafer, G.: A Mathematical Theory of Evidence. Princeton University Press, Princeton (1976)

# Nonlinear and Neural Networks Based Adaptive Control for a Wastewater Treatment Bioprocess

Emil Petre, Dan Selişteanu, Dorin Şendrescu, and Cosmin Ionete

Department of Automatic Control, University of Craiova, A.I. Cuza 13, Craiova, Romania
epetre@automation.ucv.ro

**Abstract.** The paper studies the design and analysis of some nonlinear and neural adaptive control strategies for a wastewater treatment process, which is an activated sludge process with nonlinear, time varying and not exactly known kinetics. In fact, an adaptive controller based on a dynamical neural network used as a model of the unknown plant is developed and then is compared with a classical linearizing controller. The neural controller design is achieved by using an input-output feedback linearization technique.

**Keywords:** Nonlinear systems, Neural networks, Bioprocesses.

## 1 Introduction

During the last years, the control of bioprocesses has been a significant problem attracting wide attention. The main engineering motivation in applying control methods to bioprocesses is to improve operational stability and production efficiency.

It is well known that traditional control design involves complicated mathematical analysis and has difficulties in controlling highly nonlinear and time-varying plants as well. A powerful tool for nonlinear controller design is the feedback linearization technique [1], [2], but the use of it requires the complete knowledge of the system.

Recently, there has been considerable research activity in the applications of neural networks (NN) to identification and control of nonlinear systems [3], [4]. In order to deal with the control of nonlinear uncertain systems, during the past few years, NN based adaptive control strategies have been developed [5], [6]. In [5], a combined adaptive NN scheme was designed for a class of nonlinear systems, using the sum of the output of a simple adaptive controller and the output of a NN as the control input. More recently, a NN adaptive control with input-to-state stable internal dynamics is developed in [6]. There are many application of NN adaptive control such as robotics, chemical processes etc., but a few applications for bioprocesses [7].

In this paper, the design and the analysis of some nonlinear and neural control strategies for controlling a bioprocess with incompletely known and time varying dynamics are presented. Using the feedback linearization approach, the design of a linearizing controller and of an adaptive controller based on a dynamical NN is achieved. Practically, the NN transform the originally unknown system to a dynamic NN model; the weights of this model must to be determined by adaptive techniques. The ability of dynamical NNs to learn static and dynamic highly nonlinear systems is

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 273–280, 2008.

a well-known property [8]. However, when one uses models to develop control algorithms, the presence of a modelling error term could destroy the stability of the system. In this paper the adaptive regulation problem will be developed only when the modelling error term is zero. The adaptation laws of NN weights are derived using a Lyapunov stability property of the closed loop system. The derived control method is applied in a depollution control problem, for a wastewater treatment bioprocess with strongly nonlinear, time varying and not exactly known dynamical kinetics.

## 2 Dynamical Neural Network Model and Control Strategies

Consider the class of multi-input/multi-output square nonlinear dynamical systems (that is, the systems with as many inputs as outputs) of the form:

$$\dot{x} = f(x) + \sum_{i=1}^{n} g_i(x)u_i = f(x) + G(x)u; \quad y = Cx \tag{1}$$

with the state $x \in \Re^n$, the input $u \in \Re^n$ and the output $y \in \Re^n$. $f : \Re^n \to \Re^n$ is an unknown smooth function called drift term and $G$ a matrix whose columns are the unknown smooth functions $g_i$; note that $f$ and $g_i$ contain parametric uncertainties which are not necessarily linear parameterizable. $C$ is a $n \times n$ constant matrix.

This paper deals with the control of the processes described by (1). The control objective is to make the output $y$ of (1) to track a specified trajectory $y_{ref}$. However, the problem is very difficult or even impossible to be solved if the vector fields $f$ and $g_i$ are assumed to be unknown. Therefore, in order to provide a solution to this problem, it is necessary to have a more accurate model for the unknown plant. For that purpose, in order to model the nonlinear system (1), dynamical NN are used.

*Dynamical neural networks* are recurrent, fully interconnected nets, containing dynamical elements in their neurons. They can be described by the following system of coupled first-order differential equations [8]:

$$\dot{\hat{x}}_i = a_i\hat{x}_i + b_i\sum_{j=1}^{n} w_{ij}\phi(\hat{x}_i) + b_i w_{i,n+1}\psi(\hat{x}_i)u_i, \quad i = 1,...,n \tag{2}$$

or compactly

$$\dot{\hat{x}} = A\hat{x} + BW\Phi(\hat{x}) + BW_{n+1}\Psi(\hat{x})u; \quad y_N = C\hat{x} \tag{3}$$

with the state $\hat{x} \in \Re^n$, the input $u \in \Re^n$, the output $y_N \in \Re^n$, $W$ a $n \times n$ matrix of adjustable synaptic weights, $A$ - a $n \times n$ diagonal matrix with negative eigenvalues $a_i$, $B$ - a $n \times n$ diagonal matrix of scalar elements $b_i$, and $W_{n+1}$ a $n \times n$ diagonal matrix of adjustable synaptic weights: $W_{n+1} = diag\{w_{1,n+1} \cdots w_{n,n=1}\}$. $\Phi(\hat{x})$ is a $n$-dimensional vector and $\Psi(\hat{x})$ is a $n \times n$ diagonal matrix, with elements the activation functions $\phi(\hat{x}_i)$ and $\psi(\hat{x}_i)$, usually represented by sigmoids of the form:

$$\phi(\hat{x}_i) = \frac{m_1}{1 + e^{-\delta_1\hat{x}_i}}, \quad \psi(\hat{x}_i) = \frac{m_2}{1 + e^{-\delta_2\hat{x}_i}} + \beta, \ i = 1,...,n,$$

where $m_k$ and $\delta_k$, $k = 1, 2$ are constants, and $\beta > 0$ is a constant that shifts the sigmoid, such that $\psi(\hat{x}_i) > 0$ for all $i = 1,...,n$.

Next, by using the feedback linearization technique, two nonlinear controllers for the system (1) are presented: a *linearizing feedback controller*, and a *nonlinear adaptive controller using dynamical neural networks*. Firstly, the *linearizing feedback controller* case is considered, which it is an ideal case, when maximum prior knowledge concerning the process is available. We suppose that the functions *f* and *G* in (1) are completely known, the relative degree of differential equations in (1) is equal to 1, and all states are on-line measurable. Assume that we wish to have the following first order linear stable closed loop (process + controller) behaviour:

$$(\dot{y}_{ref} - \dot{y}) + \Lambda(y_{ref} - y) = 0 \tag{4}$$

with $\Lambda = diag\{\lambda_i\}, \lambda_i > 0, i = 1,...,n$. Then, by combining (1) and (4) one obtains the following multivariable decoupling linearizing feedback control law:

$$u = \left(CG(x)\right)^{-1}\left(-Cf(x) + \nu\right) \tag{5}$$

with $(CG(x))$ assumed invertible, which applied to the process (1) result in $\dot{y} = \nu$, where $\nu$ is the new input vector designed as $\nu = \dot{y}_{ref} + \Lambda(y_{ref} - y)$. The control law (5) leads to the linear error model $\dot{e}_t = -\Lambda e_t$, where $e_t = y_{ref} - y$ is the tracking error. For $\lambda_i > 0$, the error model has an exponential stable point at $e_t = 0$.

Because the prior knowledge concerning the process is not realistic, next it will be analyzed a more realistic case, when the model (1) is practically unknown, that is the functions *f* and *G* are completely unknown and time varying. To solve the control problem, a *NN based adaptive controller* will be used. The dynamical NN (3) is used as a model of the process for the control design. Assume that the unknown process (1) can be completely described by a dynamical NN plus a modelling error term $\omega(x, u)$. In other words, there exist weight values $W^*$ and $W^*_{n+1}$ such that (1) can be written as:

$$\dot{x} = Ax + BW^*\Phi(x) + BW^*_{n+1}\Psi(x)u + \omega(x, u); \quad y = Cx. \tag{6}$$

It is clear that the tracking problem can be now analyzed for the system (6) instead of (1). Since $W^*$ and $W^*_{n+1}$ are unknown, the solution consists in designing a control law $u(W, W_{n+1}, x)$ and appropriate update laws for $W$ and $W_{n+1}$ such that the network model output $y$ tracks a reference trajectory $y_{ref}$. The dynamics of NN model output (6), where the modelling error term $\omega(x, u)$ is assumed to be 0, can be expressed as:

$$\dot{y} = C\dot{x} = CAx + CBW^*\Phi(x) + CBW^*_{n+1}\Psi(x)u \tag{7}$$

Assume that $CBW^*_{n+1}\Psi(x)$ is invertible, which implies relative degree equal to one for input-output relation (7). Then, the control law (5) is particularized as follows

$$u = \left(CBW^*_{n+1}\Psi(x)\right)^{-1}\left(-CAx - CBW^*\Phi(x) + \nu\right) \tag{8}$$

where the new input vector $\nu$ is defined as $\nu = \dot{y}_{ref} + \Lambda(y_{ref} - y)$, which applied to the model (7) results in a linear stable system with respect to this input, as $\dot{y} = \nu$.

Defining the tracking error between the reference trajectory and the network output (7), as $e_t = y_{ref} - y$, then the control law (8) leads to a linear error model $\dot{e}_t = -\Lambda e_t$. For $\lambda_i > 0, i = 1,...,n$, the error $e_t$ converges to the origin exponentially.

Note that the control input (8) is applied both to plant and neural model. Now, we can define the error between the identifier (NN) output and real system (ideal identifier) output as $e_m = y_N - y = C(\hat{x} - x)$. Assuming that the identifier states are closely to process states [8], then from (3) and (6) we obtain the next error equation:

$$\dot{e}_m = CAC^{-1}e_m + CB\tilde{W}\Phi(x) + CB\tilde{W}_{n+1}\Psi(x)u \tag{9}$$

with $\tilde{W} = W - W^*$, $\tilde{W}_{n+1} = W_{n+1} - W_{n+1}^*$. Since control law (8) contains the unknown weight matrices $W^*$ and $W_{n+1}^*$, this becomes an adaptive control law if these weight matrices are substituting by their on-line estimates calculated by appropriate updating laws. Since we are interested to obtain stable adaptive control laws, a Lyapunov synthesis method is used. Consider the following Lyapunov function candidate:

$$V = (1/2) \cdot \left( e_m^T P e_m + e_t^T \Lambda^{-1} e_t + tr\{\tilde{W}^T \tilde{W}\} + tr\{\tilde{W}_{n+1}^T \tilde{W}_{n+1}\} \right) \tag{10}$$

where $P > 0$ is chosen to satisfy the Lyapunov equation $PA + A^T P = -I$.

Differentiating (10) along the solution of (9), where $C$ is considered to be equal to identity matrix, finally one obtains:

$$\dot{V} = -1/2 \cdot e_m^T e_m - e_t^T e_t + \Phi^T(x)\tilde{W}^T BPe_m + u^T \Psi^T(x)\tilde{W}_{n+1} BPe_m + tr\{\dot{W}^T \tilde{W}\} + tr\{\dot{W}_{n+1}^T \tilde{W}_{n+1}\} \tag{11}$$

For $tr\{\dot{W}^T \tilde{W}\} = -\Phi^T(x)\tilde{W}^T BPe_m$, $tr\{\dot{W}_{n+1}^T \tilde{W}_{n+1}\} = -u^T \Psi^T(x)\tilde{W}_{n+1} BPe_m$, (11) becomes:

$$\dot{V} = -1/2 \cdot e_m^T e_m - e_t^T e_t = -1/2 \cdot \| e_m \|^2 - \| e_t \|^2 \le 0 \tag{12}$$

and consequently, for the network weights the following updating laws are obtained:

$$\dot{w}_{ij} = -b_i p_i \phi(x_j) e_{mi}, \quad i, j = 1,...,n; \qquad \dot{w}_{i,n+1} = -b_i p_i \psi(x_i) u_i e_{mi}, \quad i = 1,...,n \tag{13}$$

**Theorem 1.** Consider the control law (8), and the tracking and model errors defined above. The updating laws (13) guarantee the following properties [8]:

i) $\lim\limits_{t \to \infty} e_m(t) = 0$, $\lim\limits_{t \to \infty} e_t(t) = 0$; ii) $\lim\limits_{t \to \infty} \dot{\tilde{W}}(t) = 0$, $\lim\limits_{t \to \infty} \dot{\tilde{W}}_{n+1}(t) = 0$

*Proof.* Since $\dot{V}$ in (12) is negative semidefinite, then we have $V \in L_\infty$, which implies $e_t, e_m, \tilde{W}, \tilde{W}_{n+1} \in L_\infty$. Furthermore, $\hat{x} = x + C^{-1}e_m$ is also bounded. Since $V$ is a non-increasing function of time and bounded from below, then there exists $\lim\limits_{t \to \infty} V(t) = V(\infty)$. By integrating $\dot{V}$ from 0 to $\infty$ we obtain

$$\int_0^\infty \left( \frac{1}{2} \| e_m \|^2 + \| e_t \|^2 \right) dt = V(0) - V(\infty) < \infty \tag{14}$$

which implies $e_t, e_m \in L_2$. By definition, $\phi(x_i)$ and $\psi(x_i)$ are bounded for all $x$ and by assumption all inputs to the NN, the reference $y_{ref}$ and its time derivative are also bounded. Hence, from (8) we have that $u$ is bounded and from $\dot{e}_t = -\Lambda e_t$ and (9) we conclude that $\dot{e}_t, \dot{e}_m \in L_\infty$. Since $e_t, e_m \in L_2 \cap L_\infty$ and $\dot{e}_t, \dot{e}_m \in L_\infty$, using Barbalat's Lemma [9], one obtains that $\lim_{t\to\infty} e_t(t) = 0$ and $\lim_{t\to\infty} e_m(t) = 0$. Using now the boundedness of $u, \Phi(x), \Psi(x)$ and the convergence of $e_m(t)$ to 0, we have that $\dot{W}$ and $\dot{W}_{n+1}$ also converge to 0. But we cannot conclude anything about the convergence of weights to their optimal values. In order to guarantee this convergence, $u$, $\Phi(x)$, $\Psi(x)$ need to satisfy a persistency of excitation condition [9].

## 3 Linearizing and NN Adaptive Control of a Wastewater Process

Next, the control methods are applied for an activated sludge process, which is an aerobic process of biological wastewater treatment [10], [11]. Usually, a wastewater treatment with active sludge is operated in at least two interconnected tanks: an aerator in which the biological degradation of the pollutants takes place and a sedimentation tank (settler) in which the liquid is clarified. This process is very complex, strong nonlinear and characterized by uncertainties regarding its parameters.

In this paper a simplified model of a wastewater treatment process for the removal of two pollutants ($S_1$ and $S_2$) from the treated water will be used. The dynamics of the plant (aerator + settler) is described by the following mass balance equations [11]:

$$\begin{aligned}
&\dot{X}(t) = \mu(t)X(t) - D(t)(1+r)X(t) + rD(t)X_r(t) \\
&\dot{S}_1(t) = -(1/Y_1)\mu(t)X(t) - D(t)(1+r)S_1(t) + D(t)S_{in1} \\
&\dot{S}_2(t) = -(1/Y_2)\mu(t)X(t) - D(t)(1+r)S_2(t) + D(t)S_{in2} \\
&\dot{C}(t) = -k_c(1/Y_1 + 1/Y_2)\mu(t)X(t) - D(t)(1+r)C(t) + \alpha W(C_{\max} - C(t)) + D(t)C_{in} \\
&\dot{X}_r(t) = D(t)(1+r)X(t) - D(t)(\beta + r)X_r(t)
\end{aligned} \tag{15}$$

where $X$, $S_1$, $S_2$, $C$ and $X_r$ are the concentrations of the biomass (active sludge) in the aerated tank, of the substrate (pollutant) 1, of the substrate (pollutant) 2, of the dissolved oxygen and of recycled biomass respectively; $C_{\max}$ is the maximum concentration of dissolved oxygen, $D = F_{in}/V$ – the dilution rate ($F_{in}$ the influent flow rate and $V$ the constant aerator volume), $\mu$ – the specific growth rate, $Y_1$ and $Y_2$ are the consumption coefficients of substrates $S_1$ and $S_2$ respectively, $r$ – the rate of recycled sludge, $\beta$ – the rate of removed sludge, $k_C$ – a model constant, $W$ – the aeration rate, $S_{in1}$ and $S_{in2}$ are the concentrations of influent substrates $S_1$ and $S_2$ respectively, and $C_{in}$ is the concentration of dissolved oxygen in the inflow.

For this bioprocess, the main control objective is to maintain the wastewater degradation at a desired level despite the load and concentration variations of the pollutants. Furthermore, an adequate control of dissolved oxygen concentration in aerator is very important. Then the controlled variables are concentrations of pollutants $P = S_1 + S_2$ and dissolved oxygen $C$ inside the aerator, that is $y = [P \;\; C]^T$. Regarding the input control variables, the most realistic case is that when the rate of recycled sludge $r$ and the air flow rate $W$ are the control inputs: $u = [r \;\; W]^T$.

From (15) it can be seen that relative degrees of both controlled variables $P$ and $C$, respectively, are equal to 1; therefore a square model of the form (1) is obtained:

$$\begin{aligned} \dot{P}(t) &= -(1/Y_1 + 1/Y_2)\mu(t)X(t) - D(t)(1+r)P(t) + D(t)(S_{in1} + S_{in2}) \\ \dot{C}(t) &= -k_c(1/Y_1 + 1/Y_2)\mu(t)X(t) - D(t)(1+r)C(t) + \alpha W(C_{\max} - C(t)) + D(t)C_{in} \end{aligned} \tag{16}$$

We consider that the specific growth rate is a Monod-type model [10]:

$$\mu(t) = \mu_{\max} S_1(t) S_2(t) C(t) / [(K_{S1} + S_1(t))(K_{S2} + S_2(t))(K_C + C(t))], \tag{17}$$

with $\mu_{\max}$ – the maximum specific growth rate of microorganisms, $K_{S1}$, $K_{S2}$ – the saturation constants for the substrates, $K_C$ – the saturation constant for oxygen.

In the ideal case when the process is completely known, it can be shown, after long but direct calculations applied to linear approximations of the model (15) and (17) that, the process is minimum phase. Under these conditions, *the exactly linearizing controller* (5) with piecewise constant reference $y^* = [P^* \;\; C^*]^T$ is particularized as:

$$\begin{bmatrix} r \\ W \end{bmatrix} = B^{-1} \cdot \left( \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \cdot \begin{bmatrix} P^* - P \\ C^* - C \end{bmatrix} - \begin{bmatrix} D(S_{in1} + S_{in2} - P) \\ D(C_{in} - C) \end{bmatrix} - \begin{bmatrix} -(1/Y_1 + 1/Y_2) \\ -k_C(1/Y_1 + 1/Y_2) \end{bmatrix} \mu X \right) \tag{18}$$

where $B = \begin{bmatrix} -DP & 0 \\ -DC & \alpha(C_{\max} - C) \end{bmatrix}$ is nonsingular and so invertible as long as $P$ and $C_{\max} - C$ are different from zero (conditions satisfied in a normal process operation).

We assume now that in equation (16), the terms $(1/Y_1 + 1/Y_2)\mu(t)X(t)$ and $k_C(\mu(t)/Y_1 + \mu(t)/Y_2)X(t)$ respectively are completely unknown and time varying. Then these equations can be written in the form:

$$\begin{aligned} \dot{P}(t) &= -\eta_1 - D(t)(1+r)P(t) + D(t)(S_{in1} + S_{in2}) \\ \dot{C}(t) &= -\eta_2 - D(t)(1+r)C(t) + \alpha W(C_{\max} - C(t)) + D(t)C_{in} \end{aligned} \tag{19}$$

As in the previous case our objective is to determine the control inputs $r$ and $W$ so that $P$ and $C$ follow the desired outputs $P^*$ and $C^*$ while $\eta_1$ and $\eta_2$ are considered unknown functions. If $r$ and $W$ can be chosen as

$$\begin{bmatrix} r \\ W \end{bmatrix} = B^{-1} \cdot \left( \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \cdot \begin{bmatrix} Y^* - Y \\ C^* - C \end{bmatrix} - \begin{bmatrix} D(S_{in1} + S_{in2} - P) \\ D(C_{in} - C) \end{bmatrix} + \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} \right), \tag{20}$$

then the error equations are: $\dot{e}_1 = -\lambda_1 e_1$, $\dot{e}_2 = -\lambda_2 e_2$ with $e_1 = P^* - P$, $e_2 = C^* - C$.

Because the functions $\eta_1$ and $\eta_2$ are not known, our objective is to obtain on-line estimates of these unknown functions by using a *neural network* of the form (2):

$$\dot{\hat{\eta}}_i(t) = a_i\hat{\eta}_i + b_i\sum_{j=1}^{n} w_{ij}\phi(\hat{\eta}_j) + b_i w_{i,n+1}\psi(\hat{\eta}_i)u_i\,;\quad i,j=1,2\,. \tag{21}$$

Hence, $r$ and $W$ in (20) are modified so that the on-line estimations $\hat{\eta}_1(t)$ and $\hat{\eta}_2(t)$ are used in place of $\eta_1$ and $\eta_2$. The parameters $w_{ij}$ and $w_{i,n+1}$ are adjusted by using the adaptation laws (13).

## 4 Simulation Results and Conclusion

The performance of multivariable neural adaptive controller (20), (21), by comparison to the exactly linearizing controller (18) (which yields the best response and can be used as benchmark), has been tested by performing extensive simulation experiments.

For a proper comparison, the simulations were carried out under identical conditions. The simulations were designed so that several set point changes on the controlled variables $P$ and $C$ occurred.

The system's behaviour was analyzed assuming that the pollutant concentrations $S_{in1}$ and $S_{in2}$ act as perturbations of the form: $S_{in1}(t) = 800(1+0.2\sin(\pi t/20))$, $S_{in2}(t) = 700(1+0.2\cos(\pi t/15))$, and the kinetic coefficient $\mu_{\max}$ is time-varying as $\mu_{\max}(t) = \mu_{\max}^0(1+0.1\sin(\pi t/10))$. Also, the influent flow rate $F_{in}$ is time-varying. The gains of control laws (18), respectively (20) are $\lambda_1$ = 0.4, $\lambda_2$ = 1.5.

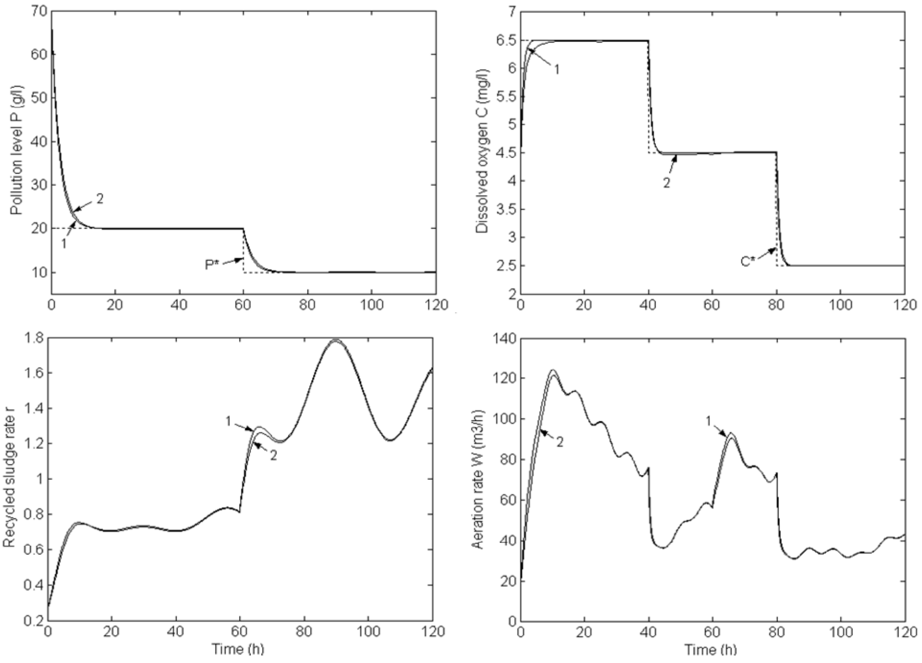


**Fig. 1.** Simulation results for neural adaptive control by comparison to exactly linearizing control: 1 – exactly linearizing control, 2 – neural adaptive control

The values of the parameters of the bioprocess are: $\mu_{\max}^{0} = 0.4h^{-1}$, $Y_1 = 0.5$, $Y_2 = 4$, $C_{in} = 8mg/l$, $k_C = 0.7$, $F_{in}^{0} = 6.5m^3/h$, $V = 100m^3$, $K_C = 2mg/l$, $K_{S1} =$ $K_{S2} = 100mg/l$, $\alpha = 0.09$, $\beta = 0.2$. The initial values of the weights are set to 0.5 and the design parameters were chosen as: $m_1 = 0.4,\ \ m_2 = 1.5,\ \ \delta_1 = \delta_2 = 0.1,$ $\beta_1 = \beta_2 = 0,\ \ a_1 = -8,\ \ a_2 = -15,\ \ b_1 = b_2 = 0.01,\ \ p_1 = p_2 = 2.5$. The behaviour of closed-loop system using NN based adaptive controller, by comparison to the exactly linearizing law is presented in Fig. 1. The first two graphics correspond to the controlled variables *P* and *C* respectively and the last two graphics correspond to the two control inputs. It can be seen that the response of the overall system with neural adaptive controller, even if this used much less priority information, is comparable to those obtained using the linearizing controller. Note also the regulation properties and the ability of NN controller to maintain the pollutant *P* at a very low level despite the high load variations (for $S_{in1}$, $S_{in2}$ and $F_{in}$), and time variation of process parameters.

It can be concluded that when process nonlinearities are not completely known and bioprocess dynamics are time varying, the NN adaptive controllers are viable alternatives. The proposed NN adaptive control strategy can be used for the control of other bioprocesses such as biomethanization process, production of some enzymes etc. [7]. As a future problem remains the controller design when the real plant is of higher order then assumed, or in the presence of other unmodelled dynamics.

## References

1. Isidori, A.: Nonlinear Control Systems - The Third Edition. Springer, Berlin (1995)
2. Sastry, S., Isidori, A.: Adaptive Control of Linearizable Systems. IEEE Trans. Autom. Control 34, 1123–1131 (1989)
3. Fidan, B., Zhang, Y., Ioannou, P.A.: Adaptive Control of a Class of Slowly Time Varying Systems with Modelling Uncertainties. IEEE Trans. Autom. Control 50, 915–920 (2005)
4. Narendra, K.S., Parthasarathy, K.: Identification and Control of Dynamical Systems using Neural Networks. IEEE Trans. Neural Networks 1, 4–27 (1990)
5. Yasser, M., Trisanto, A., Lu, J., Yahagi, T.: A Method of Simple Adaptive Control for Nonlinear Systems using Neural Networks. IEICE Transactions on Fundamentals of Electronics, Communications and Computer Science E89-A(7), 2009–2018 (2006)
6. Hayakawa, T., Haddad, W.M., Hovakimyan, N.: Neural Network Adaptive Control for a Class of Nonlinear Uncertain Dynamical Systems with Asymptotic Stability Guarantees. IEEE Transactions on Neural Networks 19, 80–89 (2008)
7. Petre, E., Selişteanu, D., Şendrescu, D.: Nonlinear and Neural Networks Based Adaptive Control for a Class of Dynamical Nonlinear Processes. WSEAS Transactions on Systems 7, 1517–1524 (2006)
8. Rovithakis, G.A., Christodoulou, M.A.: Direct Adaptive Regulation of Unknown Nonlinear Dynamical Systems via Dynamic Neural Networks. IEEE Trans. Syst. Man Cyber. 25, 1578–1594 (1995)
9. Sastry, S., Bodson, M.: Adaptive control: Stability, Convergence and Robustness. Prentice-Hall International Inc., Englewood Cliffs (1989)
10. Dochain, D., Vanrolleghem, P.: Dynamical Modelling and Estimation in Wastewater Treatment Processes. IWA Publishing, London (2001)
11. Nejjari, F., Dahhou, B., Benhammou, A., Roux, G.: Non-linear Multivariable Adaptive Control of an Activated Sludge Wastewater Treatment Process. Int. J. of Adaptive Control and Signal Processing 13, 347–365 (1999)

# Epileptic Seizure Classification Using Neural Networks with 14 Features

Rui P. Costa, Pedro Oliveira, Guilherme Rodrigues, Bruno Leitão, and António Dourado

Center for Informatics and Systems
University of Coimbra
Coimbra, Portugal
{racosta,pcoliv,gcsouza,bmleitao}@student.dei.uc.pt, dourado@dei.uc.pt

**Abstract.** Epilepsy is one of the most frequent neurological disorders. The main method used in epilepsy diagnosis is electroencephalogram (EEG) signal analysis. However this method requires a time-consuming analysis when made manually by an expert due to the length of EEG recordings. This paper proposes an automatic classification system for epilepsy based on neural networks and EEG signals. The neural networks use 14 features (extracted from EEG) in order to classify the brain state into one of four possible epileptic behaviors: inter-ictal, pre-ictal, ictal and pos-ictal. Experiments were made in a (i) single patient (ii) different patients and (ii) multiple patients, using two datasets. The classification accuracies of 6 types of neural networks architectures are compared. We concluded that with the 14 features and using the data of a single patient results in a classification accuracy of 99%, while using a network trained for multiple patients an accuracy of 98% is achieved.

**Keywords:** Neural Networks, Epilepsy, Seizure Prediction, Data Mining, Classification, EEG processing.

## 1 Introduction

Epilepsy is a common disorder that has been with us ever since ancient times, affecting about 50 million people in the world (according to the International League Against Epilepsy). Epilepsy targets the brain, a temporal change in the brains electrical activity that expresses itself in motor, psychic, sensorial and sensitive manifestations most commonly associated with spasms. Trough a visual analysis of an EEG chart a trained specialist can identify the several states of a seizure, where it begins, where it manifests onto observable characteristics and when it ends. To conduct such monitoring in real time or in a way that does not prevent the patient from performing every day tasks is a technological challenge that has the potential to minimize the impact of this illness and improve quality of life.

Techniques normally used in seizure prediction include methods based on the analysis of the EEG signal. This area of investigation generally includes,

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 281–288, 2008.

among others, the analysis of nonlinear dynamics, wavelet transform and signal quantification. In epileptic seizures prediction and detection, one of the most interesting methods is the development of computational methods described as classifiers (e.g. neural networks). The main goal of these studies is to accurately determine the epileptic EEG states through the processing of extracted EEG features. In [1] Approximated Entropy was used as feature, achieving an accuracy of 100%. However the classification was made only between two classes (normal and epileptic), this approach is not the best choice if the goal is the prediction of seizures. Another study [2] used some of the features applied in our work and obtained a classification accuracy of 96.7%. Although the dataset used were captured from a single brain region (temporal epilepsy, epileptogenic focus: hippocampal formation).

Other computational tools such as neuro-fuzzy computing techniques were recently demonstrated as highly promising in the identification of seizure patterns [3]. These efforts represent the increase variety of methods and techniques used in the processing of epileptic EEG. An extensive analysis of recent published works can be found in [4].

The number of patients analyzed in each study has a direct influence in the values of sensitivity. Usually, when the studies are based on EEG data of a small number of patients the sensitivity results tend to increase. On the other hand, when the studies reunite several patients, the sensitivity of the methods tends to decrease. This can be explained by several factors such as unique brain dynamics of each patient. Another important information is the absence in several studies of false positive rate information (specificity). The increased sensitivity cannot be obtained based in a large number of false positives. This would invalidate the development of any closed-loop seizure prevention system.

We propose several neural networks capable of classifying the different states of an epileptic seizure, using as evaluation metrics: accuracy, sensitivity and specificity. In order to build in the near future a prediction system our classification focus is on the pre-ictal state. By training different types of neural networks and testing them in different ways we attempt to identify wining characteristics that could get an accurate classification in different datasets. The classifier proposed distinguishes among four classes: inter-ictal (normal brain state), pre-ictal (just before the seizure), ictal (during seizure) and pos-ictal (after a seizure and before the normal brain state).

The organization of the paper is as follows. The next section describes the EEG Data and features extraction methods. Section 3 presents the neural networks studied including a brief description of each one. In section 4 the experiments made are described and finally in section 5 the conclusions are presented.

## 2 EEG Data

The data used in this study was collected from two patients. Both records are from the database of Freiburg Center for Data Analysis and Modeling [5]. The first one is a temporal epilepsy (with three seizures in a total of 2049 entries separated by 5 seconds) and the second a frontal epilepsy (with two seizures in a total of 1365 entries separated by 5 seconds).

The intracranial recordings utilized were acquired using Neurofile NT digital video system with 128 channels, 256 Hz sampling rate, and a 16 bit analogue-to-digital converter. Applying energy concepts, wavelet transform, nonlinear systems theory, a total of 14 features were extracted from intracranial EEG signal.

### 2.1 Features Extraction

The features extracted are listed in table 1. In [6,7] you can found a deeper explanation about these features and their extraction.

**Table 1.** Features

| Concept | Features |
|---|---|
| Signal Energy | Accumulated energy<br>Energy level<br>Energy variation (short term energy (STE))<br>Energy variation (long term energy (LTE)) |
| Wavelet Transform | Energy STE 1 (0Hz − 12.5Hz)<br>Energy STE 2 (12.5Hz − 25Hz)<br>Energy STE 3 (25Hz − 50Hz)<br>Energy STE 4 (50Hz − 100Hz)<br>Energy LTE 1 (0Hz − 12.5Hz)<br>Energy LTE 2 (12.5Hz − 25Hz)<br>Energy LTE 3 (25Hz − 50Hz)<br>Energy LTE 4 (50Hz − 100Hz) |
| Nonlinear system dynamics | Correlation dimension<br>Max Lyapunov Exponent |

**Signal energy (accumulated energy and energy variation).** Based on the algorithm presented in [8], the authors relate the EEG study with accumulated energy concept. Accumulated energy is determined by the sum of the successive values of signal energy. Then the derivative of the function is determined and analyzed, allowing the pattern evaluation; according to several authors, pre-seizure activity is related to the increase of EEG signal energy.

Accumulated energy was approximated by using moving averages of signal energy (using a short-term energy observation window vs. a long-term energy observation window).

**Wavelet transform (decomposition coefficients analysis).** The signal is decomposed in different frequency bands, and the extracted coefficients represent new functions (versions of the same original signal). The coefficients obtained by wavelet decomposition with four levels are processed and accumulated energy of these series is determined. Accumulated energy was approximated by using moving averages of coefficients energy (using a short-term energy observation window vs. a long-term energy observation window).

**Nonlinear dynamics.** Several approaches, based on the chaos theory, were used successfully in EEG analysis; due to the aperiodic and instable behavior of the epileptic brain, the structure is suitable to nonlinear techniques. Functions designed for this purpose (TSTOOL[9] matlab toolbox) were used to process EEG signal and determine the Lyapunov exponents (quantification of the exponential growth of the average distance between two nearby trajectories through error approximation) and correlation dimension (estimator method) of signal short segments.

### 2.2 Feature Preparation

The datasets were normalized by feature in the interval [0 1]. This normalization gives an identical influence of each feature for the calculation of neural network weights.

## 3 Neural Networks Applied

After some preliminary tests, we chose six neural network variants to our study. These neural networks cover a wide spectrum of the available neural network approaches, allowing us to gather a good knowledge about the use of neural networks in the Epileptic Seizure Detection problem. For a more comprehensive explanation about the neural networks described in the next subsections see [10].

### 3.1 Radial Basis Function (RBF)

In our study we used Exact Fit variant [11], where the number of neurons in the 1st layer is equal to the number of prototypes in the input (in our case, 14). The spread constant used, i.e. the area of input space to which each neuron responds, was 1.5.

### 3.2 Feed-Forward BackPropagation (FFBP) and Layer-Recurrent Networks (LRN)

FFBP [12] and LRN [13] have been used. LRN are composed by an arbitrary number of layers, with a feedback loop around each layer, except for the output layer. This feedback loop provides a single delay to the network. Our networks were configured using 2 layers, being the hidden layer composed by 10 tansig neurons and the output layer by 4 linear neurons. Both networks were trained with the Levenberg-Marquardt algorithm.

### 3.3 Elman and Distributed Time Delay (DTD) Networks

We used Elman Networks [13] with one hidden layer, composed by 10 tansig neurons, followed by a linear output layer, with Lvenberg-Marquardt backpropagation function. Distributed Time Delay networks [14] are dynamic neural networks, where the output of the various layers also depends on the past output of these layers. This capability is achieved by using tapped delay line memories, which record the past outputs of each layer. Our network was designed with a hidden layer of 10 tansig neurons, with a boolean output layer. The training function was the Levenberg-Marquardt backpropagation function, and we used a one step time-delay.

### 3.4 Feed-Forward Input Time-Delay BackPropagation (FFTD)

Input Time-Delay networks [14] are very similar to Feed-Forward Networks trained with the backpropagation algorithm. The only difference is that they take as inputs not only the training data, but also a predefined time-delay from the data. Therefore, they can deal with temporal and spatial data. Our configuration is very similar to Feed-Forward BackPropagation (one hidden layer with 10 tansig neurons, and Levenberg-Marquardt backpropagation training function). The Input Time-Delay used was one time unit.

## 4 Experiments

In order to apply the neural networks previously presented we used Matlab R2007b with Neural Networks Toolbox. Within this platform we implemented several scripts that allowed us to run the experiments. The developed code and the data files for training and testing are available[1].

### 4.1 Evaluation Metrics

To evaluate our results we used three different metrics: *sensitivity* (1), i.e. the capacity of correctly identify positive cases (pre-ictal), *specificity* (2), i.e. the capacity of correctly identify negative cases (non pre-ictal), and *accuracy* (3), i.e. the proportion of correct classified instances. The use of four brain states is useful in order to better evaluate the accuracy of the classifiers built. These metrics are largely used in this domain, making easier to compare our results with other works. In order to implement these metrics each entry of the datasets were previously classified by a medical expert as: inter-ictal, pre-ictal, ictal or pos-ictal.

$$Specificity(\%) = \frac{True\ \ Negatives}{True\ \ Negatives + False\ \ Positives} \times 100 \tag{1}$$

$$Sensitivity(\%) = \frac{True\ \ Positives}{True\ \ Positives + False\ \ Negatives} \times 100 \tag{2}$$

$$Accuracy(\%) = \frac{Correct\ \ cases}{Total} \times 100 \tag{3}$$

[1] http://student.dei.uc.pt/˜racosta/epilepsy

**Table 2.** Results of the several experiments SP - Specificity, SS - Sensitivity, AC - Accuracy

| | RBF | | | FFBP | | | Elman | | | Recurrent | | | FFTD | | | DTD | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | SP | SS | AC | SP | SS | AC | SP | SS | AC | SP | SS | AC | SP | SS | AC | SP | SS | AC |
| Single(1) | 96 | 98 | 93 | **99** | **98** | **98** | **99** | **93** | **98** | **99** | **97** | **97** | **99** | **97** | **98** | **99** | **97** | **98** |
| Single(2) | 97 | 97 | 91 | **100** | **100** | **99** | **99** | **100** | **98** | **100** | **100** | **98** | **99** | **100** | **98** | **99** | **97** | **98** |
| Different(1:2) | 89 | 2 | 63 | 66 | 26 | 54 | 85 | 0 | 8 | 32 | 94 | 28 | 74 | 60 | 64 | 81 | 0 | 22 |
| Different(2:1) | 93 | 0 | 2 | 77 | 3 | 44 | 99 | 0 | 48 | 59 | 53 | 18 | 96 | 0 | 68 | 85 | 34 | 42 |
| Multiple(1+2:1) | **100** | **97** | **100** | 99 | 99 | 98 | **99** | **92** | **98** | 98 | 76 | 95 | 99 | 65 | 96 | **99** | **96** | **98** |
| Multiple(1+2:2) | **100** | **100** | **99** | 99 | 82 | 97 | **98** | **96** | **97** | 100 | 41 | 91 | 96 | 92 | 93 | **99** | **90** | **97** |

### 4.2 Single Patient

In this test, we used the data extracted from a single patient to test and train the neural networks. Were used 70% of the patient data to train the network, while the others 30% were used to test it. From each set of 3 entries were taken 2 entries to the training set and the other one to the testing set. The results obtained using the neural networks in two different patients are shown in Table 2 (row 1 for patient A and row 2 for patient B).

We can see some interesting results for the three performance criteria (1),(2) and (3). FFBP, Elman, Recurrent, FFTD and DTD show very good results in both patients.

### 4.3 Different Patients

In this test, we trained the network with the data from one patient and tested it with the other patient. The results obtained are shown in Table 2 (rows 3 and 4).

There is an evident degradation of performance. This is probably because the patients have different kinds of epilepsy and the networks do not have sufficient generalization capability. This seems to indicate that seizure prediction with neural networks needs a personalized network, specific for each patient.

### 4.4 Multiple Patients

In this test, we trained the network with the data from both patients, and tested with one of them. This was done by concatenating both datasets into one. The results obtained are shown in Table 2 (rows 5 and 6).

In this case there are still good results, with Elman and DTD networks, both with memory. This is an interesting indication that the memory may improve the generalization capability of the network. The RBF neural network obtained very good results, however the networks had more than 3000 neurons. This can lead to the impossibility of training these networks due to their excessive memory needs, at the same time this seems to show that they are strongly addicted to the training datasets (almost one neuron for each data entry). In the future the use of more than two datasets to train the neural networks should be studied.

## 5 Conclusions and Discussion

In this paper we propose the classification of epileptic EEG data into four states (inter-ictal, pre-ictal, ictal and pos-ictal) applying several neural networks architectures.

From the EEG data were extracted 14 features: accumulated energy, level, lyapunov exponents, correlation dimension, five variants of energy STE and five variants of energy LTE.

The classification accuracies of the following neural networks are compared: 1) Radial basis function, 2) Feed-Forward BackPropagation, 3) Layer-Recurrent, 4) Elman, 5) Feed-Forward Input Time-Delay BackPropagation, 6) Distributed Time Delay.

The results show that it is possible to find a good classifier to the four brain states based on neural networks (with an accuracy of 99%). However the classifier of one patient cannot be used for another patient. The variability of physiological systems can only be overcome personalizing the architecture and the training of the network. The performance of the classifier, if it is intended to be used for example to give the patient an alarm of an approaching seizure (classifying correctly the pre-ictal state), must be checked by both sensitivity and specificity. If one limits to only one of them, no practical usefulness can be given to the results.

Considering the brain as a nonlinear complex dynamic system, memory in the networks seems to be a natural element. Further research is needed to find more elaborated memory architectures and its appropriate training algorithms. Neural networks as classifiers have here a high potential because they can compute in real time with a high number of features. This characteristic enable the development and construction of transportable devices, improving substantially the quality of life of epileptic patients intractable by medication and that must learn to live with seizures.

## Acknowledgments

This research is supported by European Union FP7 Framework Program, EPILEPSIAE Grant 211713 and by CISUC (FCT Unit 326). The authors express also their gratitude to Freiburg Center for Data Analysis and Modeling (FDM) of Albert Ludwigs University of Freiburg, for the access to the epilepsy database.

## References

1. Srinivasan, V., Eswaran, C., Sriraam, N.: Approximate entropy-based epileptic EEG detection using artificial neural networks. IEEE Transactions On Information Technology In Biomedicine 11(3), 288–295 (2007)
2. Ghosh-Dastidar, S., Adeli, H., Dadmehr, N.: Mixed-band wavelet-chaos-neural network methodology for epilepsy and epileptic seizure detection. IEEE Transactions on Biomedical Engineering 54(7), 1545–1551 (2007)

3. Subasi, A.: Application of adaptive neuro-fuzzy inference system for epileptic seizure detection using wavelet feature extraction. Comput. Biol. Med. 37(2), 227–244 (2007)
4. Mormann, F., Andrzejak, R.G., Elger, C.E., Lehnertz, K.: Seizure prediction: the long and winding road. Brain 130, 314–333 (2007)
5. Freiburger Zentrum fur Datenanalyse und Mollbildung: The freiburg seizure prediction project database, `https://epilepsy.uni-freiburg.de/freiburg-seizure-prediction-project/eeg-database`
6. Leitão, B., Dourado, A., Vieira, M., Sales, F.: Computational system for the prediction of epileptic seizures through multi-sensorial information analysis. Technical report, Department of Informatics Engineering, University of Coimbra (September 2007)
7. Winterhalder, M., Schelter, B., Maiwald, T., Brandt, A., Schad, A., Schulze-Bonhage, A., Timmer, J.: Spatio-temporal patient-individual assessment of synchronization changes for epileptic seizure prediction. Clinical Neurophysiology 117, 2399–2413 (2006)
8. Litt, B., Esteller, R., Echauz, J., D'Alessandro, M., Shor, R., Henry, T., Pennell, P., Epstein, C., Bakay, R., Dichter, M., Vachtsevanos, G.: Epileptic seizures may begin hours in advance of clinical onset: A report of five patients. Neuron. 30(1), 51–64 (2001)
9. Merkwirth, C., Parlitz, U., Wedekind, I., Lauterborn, W.: Tstool user manual, version 1.11 (2001), `http://www.dpi.physik.uni-goettingen.de/tstool/HTML/index.html`
10. Demuth, H., Beale, M., Hagan, M.: Neural network toolbox 6 user's guide. The MathWorks (2008)
11. Chen, S., Cowan, C., Grant, P.M.: Orthogonal least squares learning algorithm for radial basis function networks. IEEE Transactions on Neural Networks 2(2), 302–309 (1991)
12. Hagan, M.T., Demuth, H.B., Beale, M.: Neural Network Design. PWS Publishing Company (1996)
13. Elman, J.L.: Finding structure in time. Cognitive Science 14, 179–211 (1990)
14. Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., Lang, K.: Phoneme recognition using time-delay neural networks. IEEE Transactions on Acoustics, Speech and Signal Processing 37(3), 328–339 (1989)

# Predicting Grid Performance Based on Novel Reduct Algorithm*

Kun Gao

Computer Science and Information Technology College, Zhejiang Wanli University
No. 8, South Qian Hu Road, 315100, Ningbo, Zhejiang, China
gaoyibo@gmail.com

**Abstract.** Because of the irregular characteristic of Grid environment, we are unable to predict the performance using the traditional method. In this paper, we propose a novel method for predicting the performance in Grid Computing environment. The method, based on frequencies of application attributes appeared in discernibility matrix collected during a period of time; predict the applications performance that the traditional methods can't obtain. We use the novel method in Data Ming Grid and obtain better result than traditional methods. The results of the experiment show that the use of reduct algorithm can process uncertain problem in Data Mining Grid. The theoretical foundation of ruduct provides a feasible solution to the problem of predicting Data Mining Grid.

**Keywords:** Predicting performance, Distributed computing, Soft computing.

## 1 Introduction

Grid computing [1] [2] [3] is driven by the vast improvements in wide area network performance. It has emerged as a promising attractive computing paradigm. Computational Grids aim to aggregate the power of heterogeneous, geographically distributed, multiple domain spanning computational resources to provide high performance or high throughput computing. To achieve the promising potentials of computational Grids, an effective and efficient scheduling system is fundamentally important. In concept, Grid computing is a subset of distributed computing; in function, Grid computing is expansion and continuity to distributed computing. Grid emphasizes coordination and cooperation between Grid resources.

It becomes the encouraging trend, because of the following reasons:

1. Grid computing can effectively make use of the existing resources.
2. It can condense a large amount of computing capability to solve the problem which can not be solved before grid.
3. It will build widely distributed computing platform to integrate all kinds of resource including computation power resource, data resource, network resource and so on.

In this paper, we propose a novel method for predicting the performance in Grid Computing environment. We use the concept of Reduct in Rough Set theory and

[*] Programs Supported by Ningbo Natural Science Foundation (2008A610028).

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 289–296, 2008.

history record collected during a period of time to predict the applications runtime that the traditional methods can't obtain. The approach is based on frequencies of attributes appeared in discernibility matrix. The theoretical foundation of rough sets provides an intuitive solution to the problem of application run time estimation on Data Ming Grid. The results of the experiment show that the use of Rough Set theory can process uncertain problem in distributed and dynamic environment, and obtain better result than traditional methods.

The rest of this paper is organized as followed: We introduce some related works and related reduct concept in section 2, 3; and then we propose a novel reduct algorithm in section 4; in section 5, we introduce the data mining framework, DMG. We conduct experiment to evaluate our approach in section 6. Finally in section 7, we conclude this paper.

## 2 Related Works

Early work in the parallel computing area proposed using similarity templates of application characteristics to identify similar tasks in a history [4] [5] [6] [7]. Recently, another effective approach to predict execution times on Grid is [8]. They investigate a use of sampling: in order to forecast the actual execution times of a given data mining algorithm on the whole dataset, they run the same algorithm on a small sample of the dataset. Many data mining algorithms demonstrate optimal scalability with respect to the size of the processed dataset, thus making the performance estimate possible and accurate enough. However, in order to derive an accurate performance model for a given algorithm, it should be important to perform an *off-line* training of the model, for different dataset characteristics and different parameter sets.

Presented in [9] is a method, named the Dynamic Rule Prediction (DRP), which predicts the behavior of a system and its application in designing a controller. The aim of the author is to overcome some of the limitations and shortcoming of the other modeling methods. The effectiveness of this method is verified. The controller based on DRP possesses two main features. It can control the system without any prior knowledge of the controlled plant. It is, also, superior as its high-speed prediction. The author focuses on the robot manipulator controllers and applications of this approach in it.

In designing running mechanism of a tracked vehicle running on off-road, the sinkage of the vehicle should be predicted by some methods to optimize the running mechanism. In order to verify the accuracy of these empirical models, seven plates with similar load and contact area but different dimensions of length and width were employed as well as a new pressure-sinkage measurement system equipped by a manipulator. The results of the tests show that these empirical models are not always coincident with the real situation. So [10] proposed a new method to predict the sinkage of a plat more accurately, by using fuzzy neural network. Because the soil parameters and the load exerted on the plat are ambiguous, we propose a fuzzy prediction method using possibility function to express the distribution of the sinkage.

In [11], author presents the possibility of the design of frontal neural networks and feed-forward neural networks (without pre-processing of inputs time series) with learning algorithms on the basis genetic and eugenic algorithms and Takagi-Sugeno fuzzy inference system (with pre-processing of inputs time series) in predicting of gross domestic product development by designing a prediction models whose accuracy is superior to the models used in praxis.

In this paper, we develop a rough sets based technique to address the problem of *automatically* selecting characteristics that best define similarity. In contrast to above methods, our method determines a reduct as template, instead of using greedy and genetic algorithms. Rough sets provide an intuitively appropriate theory for identifying templates. The entire process of identifying similarity templates and matching tasks to similar tasks is based on rough sets theory, thereby providing an appropriate solution with a strong mathematical underpinning.

## 3 Reduct Concept

An attribute $a$ is dispensable in $B \subseteq A$ if $POS_B(d)= POS_{B-\{a\}}(d)$. A reduct of $B$ is a set of attributes $B' \subseteq B$ such that all attributes $a \in B-B'$ are dispensable, and $POS_B(d)= POS_{B'}(d)$.

A reduct consists of the minimal set of condition attributes that have the same discerning ability as the original IS. In other words, the reduct includes the most significant attributes. All reducts of a dataset can be found by constructing a kind of discernibility function from the dataset and simplifying it. Unfortunately, it has been shown that finding minimal reduct or all reducts are both NP-hard problems.

There are usually many reducts in an information system. In fact, one can show that the number of reducts of an information system may be up to $C^{|A|/2}_{|A|}$. In order to find reducts, discernibility matrix and discernibility function are introduced.

The discernibility matrix of an information system is a symmetric matrix:

$|U|\times|U|$

with entries $c_{ij}$ defined as:

$\{a \in A|a(x_i)\neq a(x_j)\}$ if $d(x_i)\neq d(x_j)$, $\Phi$ otherwise.

A discernibility function can be constructed from discernibility matrix by or-ing all attributes in $c_{ij}$ and then and-ing all of them together. After simplifying the discernibility function using absorption rule, the set of all prime implicants decides the set of all reducts of the IS.

## 4 A Novel Reduct Algorithm

The heuristic comes from the fact that intersection of a reduct and every items of discernibility matrix can not be empty. If there are any empty intersections between some item $c_{ij}$ with some reduct, object i and object j would be indiscernible to the reduct. And this contradicts the definition that reduct is the minimal attribute set discerning all objects (assuming the dataset is consistent).

A straightforward algorithm can be constructed based on the heuristic. Let candidate reduct set $R=\Phi$. We examine every entry $c_{ij}$ of discernibility matrix. If their intersection is empty, a random attribute from $c_{ij}$ is picked and inserted in $R$; skip the entry otherwise. Repeat the procedure until all entries of discernibility matrix are examined. We get the reduct in R.

The algorithm is simple and straightforward. However, in most times what we get is not reduct itself but superset of reduct. For example, there are three entries in the matrix: $\{a_1, a_3\}$, $\{a_2, a_3\}$, $\{a_3\}$. According the algorithm, we get the reduct $\{a_1, a_2, a_3\}$ although it is obvious $\{a_3\}$ is the only reduct. This is because that our heuristic is a necessary but not sufficient condition for a reduct. The reduct must be a minimal one. The above algorithm does not consider this. In order to find reduct, especially shorter reduct in most times, we need more heuristics.

A simple yet powerful method is sort the discernibility matrix according $|c_{ij}|$. As we know, if there is only one element in $c_{ij}$, it must be a member of reduct. We can image that attributes in shorter and frequent $|c_{ij}|$ contribute more classification power to the reduct. After sorting, we can first pick up more powerful attributes, avoid situations like example mentioned above, and more likely get optimal or sub-optimal reduct.

The sort procedure is like this. First, all the same entries in discernibility matrix are merged and their frequency is recorded. Then the matrix is sorted according to the length of every entry. If two entries have the same length, more frequent entry takes precedence.

When generating the discernibility matrix, frequency of every individual attribute is also counted for later use. The frequencies is used in helping picking up attribute when it is need to pick up one attribute from some entry to insert into reduct. The idea is that more frequent attribute is more likely the member of reduct. The counting process is weighted. Similarly, attributes appeared in shorter entry get higher weight. When a new entry $c$ is computed, the frequency of corresponding attribute $f(a)$ are updated as $f(a)=f(a)+|A|/|c|$, for every $a \in c$; where $|A|$ is total attribute of information system. For example, let $f(a_1)=3, f(a_3)=4,$ the system have 10 attributes in total, and the new entry is $\{a_1, a_3\}$. Then frequencies after this entry can be computed: $f(a_1)=3+10/2=8; f(a_3)=4+10/2=9.$

## 5 The Structure of the DMG

The DMG, Data Mining Grid, is a dynamic and distributed environment where data mining application is running. Its core component is task scheduling and resource allocation. These key questions can be solved through the algorithm proposed by us.

The Data Mining Grid is mainly made up by following components:

### 5.1 DMGrid Client Node

In consideration of ease of use, the system adopts Browser/Server mode. Grid client exchanges information with Grid portal through Internet Explorer browser. Users submit the requirement of data mining and receive the final result at Grid client.

### 5.2 DMGrid Portal Node

It provides a single access way to distributed data mining application based grid. Users can make use of the whole grid resource transparently through the grid portal. This component is responsible for translating users' demand into the RSL language (Resource Specification Language) that can be recognized by grid, is used for grid resource discovery and grid resource allocation management. The final result is returned to grid portal first, and then returned to users by the portal.

### 5.3 DMGrid Resource Broker Node and DMGrid Tasks Allocation Broker Node

User's data mining requirement has driven grid resource discovery. According to users' demand condition, DMGrid resource broker looks for the resources which meet the condition in a large number of grid resources, including algorithms, computing capability and data resource. It is an important job that finds appropriate resource [12] [13]. As to any application based on grid, it is first to find appropriate resource, then allocate tasks and management them. It can be predicted that there may be many nodes which meet a condition. Resource broker is used for finding available resource among MDS (Meta Directory Service); mapping between data resource and computing resource, i.e., the task allocation broker is responsible for dispatching a certain task on a certain node.

### 5.4 Grid Node

The Grid nodes are made up of personal computer, high performance computer and cluster. Each node is installed GLOBUS, as grid middleware. They are the data carrier and the computation implementation entity.

The rationale of design and development the grid enabled data mining system is as follows:

- DMGrid adopts the standard, common and open grid service mode, follows OGSA norm, and offers unified support to the data mining applications.
- Based on Globus Toolkit and according to the existing networks system structure, DMGrid use the grid service to realize communication, operates each other and resource management.
- DMGrid is open, supports various data mining tools and algorithms, the extensibility is good.
- DMGrid is able to realize the improvement of performance by increasing network node, high performance computing node and cluster, the scalability is strong.
- DMGrid can deal with distributed huge volumes of high dimensional dataset, support heterogeneity data source.
- The main purpose to design and develop the DMGrid system is to improve the performance.
- Users carry out the data mining tasks in a transparent way; the concrete system structure, operation and characteristic in the grid environment is to be hidden.
- In the field of data mining, the security of the data and personal secrets are a sensitive topic. Data Ming Grid supports the choice of place that the data mining execute.

## 6 Primary Result of Experiment

We contact the experiment in the Data Mining Grid, The simulated environment is composed of three machines which installed with GT3[14]. Each machine is

**Table 1.** Condition Attributes and Corresponding Reduct in Each Experiment

| Experiment Number | Condition Attributes | Reduct |
|---|---|---|
| 1 | time, algorithm, parameter, disk cache, data size | algorithm, parameter, data size |
| 2 | operating system, time, algorithm, parameter, disk cache, data size, dimensionality, file name | algorithm, parameter, data size, dimensionality |
| 3 | CPU type, operating system, time, algorithm, parameter, disk cache, data size, dimensionality, file name, operating system version, CPU speed | algorithm, parameter, data size, dimensionality, CPU speed |
| 4 | memory type, CPU type, operating system, time, algorithm, parameter, disk cache, data size, dimensionality, file name, operating system version, CPU speed, available memory, disk type | algorithm, parameter, data size, dimensionality, CPU speed, available memory |
| 5 | IP, memory type, CPU type, operating system, time, algorithm, parameter, disk cache, data size, dimensionality, file name, operating system version, CPU speed, available memory, disk type, bandwidth, mainboard bus | algorithm, parameter, data size, dimensionality, CPU speed, available memory, bandwidth |

interconnected by a switched fast Ethernet. Three distributed machines with different physical configurations and operating systems: a Pentium III running Windows 2000 with an 833-MHz processor and 512 Mbytes of memory; a Pentium 4 running Windows 2000 with a 2.0 GHz processor and 1Gbytes of memory; and a Sun Sparc station running Sun OS 5.8 with a 444-Mhz processor and 256 Mbytes of memory. For each data-mining job, we recorded the following information in the history: the algorithm, file name, file size, operating system, operating system version, IP address of the local host on which the job was run, processor speed, amount of memory, bandwidth, and start and end times. We used histories with 100 and 150 records, and as before, each experimental run consisted of 25 tests.

We differentiated the test case from the historical records by removing the runtime information. Thus, a test case consists of all the information specified except the recorded runtime. The runtime information recorded in the test case was the task's actual runtime. The idea was to determine an estimated runtime using our prediction technique and compare it with the task's actual runtime.

We compiled a history of data-mining tasks by running several data-mining algorithms and recording information about the tasks and environment. We executed several runs of data-mining jobs by varying the jobs' parameters such as the mining algorithm, the data sets and the sizes of the data sets. The algorithms we used were from the Weka package of data-mining algorithms. We generated several data sets of sizes varying from 1 to 20 Mbytes.

In our experiment, the mean error was 0.23 minutes, and the mean error as a percentage of the actual runtimes was 7.6 percent. This shows that we accurately estimated the runtime for data-mining tasks on Grid. The reduct that our algorithm selected as a similarity template included the bandwidth, algorithm, file size, dimensionality, and available memory attribute. Table 1 shows the condition attributes and corresponding reduct in each experiment. Because the reduct algorithm operate entirely on the basis of the condition attributes available in the history and require no external additional information, thus the more abundant the information correlating with performance, the more accurate the prediction is.

## 7 Conclusions and Future Works

We have presented a novel reduct to predicting the performance of Data Ming Grid. The approach is based on frequencies of attributes appeared in discernibility matrix. The theoretical foundation of reduct provides an intuitive solution to the problem of application performance estimation on Data Mining Grid. Our hypothesis that reduct algorithm is suitable for predicting performance in Data Mining Grid is validated by the experimental results, which demonstrate the good prediction accuracy of our approach.

In the future, we will use the technique to build a wide Web service interface, so that some applications can visit this interface to obtain the performance parameter which they needed for enhancing the performance of system.

## References

[1] Cannataro, M., Talia, D., Trunfio, P.: KNOWLEDGE GRID: High Performance Knowledge Discovery Services on the Grid. In: Lee, C.A. (ed.) GRID 2001. LNCS, vol. 2242. Springer, Heidelberg (2001)

[2] Foster, I., Kesselman, C. (eds.): The Grid: Blueprint for a Future Computing Inf. Morgan Kaufmann Publishers, San Francisco (1999)

[3] Chervenak, A., Foster, I., Kesselman, C., Salisbury, C., Tuecke, S.: The Data Grid: towards an architecture for the distributed management and analysis of large scientific datasets. J. of Network and Comp. Appl. (2001)

[4] Downey, A.B.: Predicting Queue Times on Space-Sharing Parallel Computers. In: Proc. 11th Int'l ParallelProcessing Symp. (IPPS 1997). IEEE CS Press, Los Alamitos (1997)

[5] Gibbons, R.: A Historical Application Profiler for Use by Parallel Schedulers. In: Feitelson, D.G., Rudolph, L. (eds.) IPPS-WS 1997 and JSSPP 1997. LNCS, vol. 1291. Springer, Heidelberg (1997)

[6] Smith, W., Taylor, V., Foster, I.: Predicting Application Runtimes Using Historical Information. In: Feitelson, D.G., Rudolph, L. (eds.) IPPS-WS 1998, SPDP-WS 1998, and JSSPP 1998. LNCS, vol. 1459, pp. 122–142. Springer, Heidelberg (1998)

[7] Smith, W., Taylor, V., Foster, I.: Using Runtime Predictions to Estimate Queue Wait Times and Improve Scheduler Performance. In: Feitelson, D.G., Rudolph, L. (eds.) JSSPP 1999, IPPS-WS 1999, and SPDP-WS 1999. LNCS, vol. 1659, pp. 202–229. Springer, Heidelberg (1999)

[8] Orlando, S., Palmerini, P., Perego, R., Silvestri, F.: Scheduling high performance data mining tasks on a data grid environment. In: Proceedings of Europar (2002)

[9] Damangir, S., Jafarijashemi, G., Zohoor, H.: Modelling of a complex system using the dynamic rule prediction. WSEAS Transactions on Systems 5(12), 2833–2838 (2006)

[10] Saulia, L., Chen, P., Koji, N.: Intelligent method of sinkage prediction for tracked vehicles using possibility theory and fuzzy neural network. WSEAS Transactions on Systems 6(6), 1110–1115 (2007)

[11] Olej, V.: Design of the models of neural networks and the Takagi-Sugeno fuzzy inference system for prediction of the gross domestic product development. WSEAS Transactions on Systems 4(4), 314–319 (2005)

[12] Allen, G., Benger, W., Goodale, T., Hege, H., Lanfermann, G., Merzky, A., Radke, T., Seidel, E., Shalf, J.: The Cactus Code: A Problem Solving Environment for the Grid. In: Proceedings of the Ninth International Symposium on High Performance Distributed Computing (HPDC). IEEE Press, Pittsburgh

[13] Marzullo, K., Ogg, M., Ricciardi, A., Amoroso, A., Calkins, F., Rothfus, E.: NILE: Wide-Area Computing for High Energy Physics. In: Proceedings of 7th ACM SIGOPS European Workshop, Connemara, Ireland, September 2-4. ACM Press, New York (1996)

[14] Globus Toolkit, `http://www.globus.org/ogsa/releases/alpha/`

# Discussing Redundancy Issues in Intelligent Agent-Based Non-traditional Grids

Versavia Ancusa, Razvan Bogdan, and Mircea Vladutiu

"Politehnica" University of Timisoara, V. Parvan no. 2,
300223, Timisoara, Romania
{vancusa, rbogdan, mvlad}@cs.utt.ro

**Abstract.** In this paper is presented an improved model of ambient intelligent based on non-traditional Grids. It is studied the fault tolerance of the proposed model taking into account the redundancy at link level.

## 1 Introduction

A new paradigm – shifting technology [1] started to emerge in the late years, as a third wave of computing and it represents a new way for the interaction between the electronics and the human individuals. The third wave brought the collection of devices and their reactive interface into focus. A concept included in this third wave is "ambient intelligence" being described [8] as a sensitive, adaptive, and responsive to the presence of people and objects environment where technology is embedded, hidden in the background and augments activities through smart non-explicit assistance. This environment should also preserve the security, privacy and trustworthiness while utilizing information when needed and appropriate. Ambient intelligence is designed to be used into an environment suitable to be controlled; therefore its primary target is the "smart" home [7]. Into such an environment, besides the sensor, actuators and multimedia processing another participant might appear: the PCs. This participant is not taken into consideration in the ambient intelligence approach. On the other hand, today the multimedia ambient concept is centered on a PC approach (e.g. Microsoft Media Center). This PC controls the network. Providing the case this center fails, the whole network is down. The user is the system and configuration manager making this concept still connection and device oriented.

The idea behind our approach is that the PC is just another node in the network. It can move (spatially), it can disappear or appear at will, or it can sleep. When it is awake and the human user interacts with it, it can be seen as a merged sensor-actuator entity. When it is awake it provides computational power, therefore the network can use it to process operations. Doing so, other nodes in the network can sleep, therefore reducing overall power consumption. Any concept involving ambient intelligence cannot be addressed without taking into account the sensors within it. One of the major challenges in building such networks is to maintain long network lifetime but also sufficient sensing area. To achieve this goal, a broadly-used method is to turn off redundant sensors. This article addresses the problem of redundancy among sensors in

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 297–305, 2008.

a third wave architecture. The results presented here can be employed in designing effective sensor scheduling algorithms to reduce energy consumption while maintaining a fault-tolerant architecture.

## 2 Non-traditional Parallel Model

In order to entirely describe the intelligent grid it should be stated that, while ambient intelligence represents a composition of the sensor and multimedia networks, the intelligent grid represents a composition of the sensor and multimedia networks and computational network, as well.

The levels of this proposed model are illustrated in figure 1. The pieces of equipment (sensors, actuators, controllers, PCs, DSPs) though extremely heterogeneous, are all grouped into the device level. The links between them, as well as the association to the Internet are at the connection level, while the collection level is composed not only of streaming data sources but also of all the networks and their capabilities. In the same figure, the individual networks are also denoted.

Studying this approach from the power, cost and size style, due to its heterogeneous nature, this model must be assessed by device. The sensors, actuators, controllers and DSPs are already very well on the way of becoming "disappearing electronics". Many solutions [11] emerged in the last years, though lately the idea of energy harvesting [12] is looked at from a system approach. All this is encouraged by the results of the industry [13]. The main problem in "disappearing electronics" is that

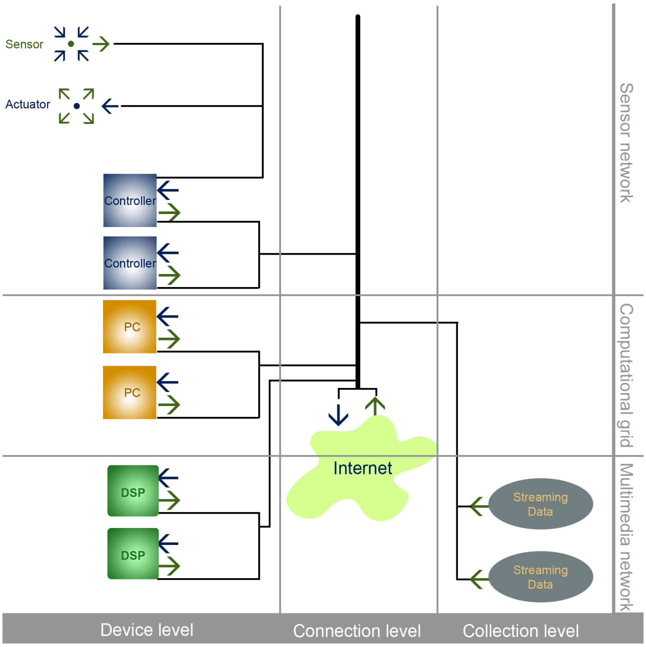


**Fig. 1.** The intelligent grid

these super low-power devices should need no batteries or outside power supply, instead relying upon microgenerators. Thus, the area of nanogenerators is blossoming. Two main areas are developing in this area: nanoscale thermoelectric energy harvesting [14] and Nano-Piezotronics[15, 16]. The PCs are forecasted to disappear for a while now [17], and every major company [18] is preparing for it.

Security and privacy [19, 20] is a thorny problem in this type of networks, due to its inherent mix and distributed nature. Portability, scalability and configurability are another problem. Changing of the application usually means changing the software, which, of course, implies everything to fall apart. A real solution is considered to be the raising of the abstraction level [10].

Reliability [30] is another major problem, due to the fact that unreliability is inherent to the disappearing electronics concept. This is primary caused by the fact that nodes may emerge unexpectedly, may move, may fail and may finish their energy reserves (temporary or not). All this problems are forced into focus by the cost, power and size constraints. The solution used in ambient intelligence in order to achieve reliability is redundancy [10, 7]; therefore we expect it to be also a solution.

### 2.1 A Middleware Discussion

During the programming of any networks, first of all is a significant need for programming abstractions that simplify tasking, and for a middleware that supports such programming abstractions [21]. However, we must take into account the fact that our grid is composed of three distinct types of network, with their specific requirements. A summary of the programming requirements for each individual type of network is presented in table 1.

**Table 1.** Overview of the programming requirements

| Programming requirements | | Networks based on | | |
|---|---|---|---|---|
| | | Sensor | Multimedia | Computation |
| Concealed issues | | hardware and distribution | hardware | distribution |
| Restricted Resources | | | | |
| | Energy | X | | |
| | Computing power | X | X | X |
| | Bandwidth | X | X | X |
| Network Dynamics | | high | medium | low |
| Scale of Deployments | | N*(100…1000) | N*(10…100) | N*(10…100) |
| Real-world Integration | | | | |
| | Time scale | X | X | X |
| | Location scale | X | X | |
| Collection and Processing of Data | | | | |
| | Preprocessing | X | X | |
| | Aggregating data | X | X | X |
| | Local processing | | X | X |

In choosing a middleware we must see if it can support all the requirements in their worse case scenario.

There are already a series of middleware for every type of network, and in table 2 we present the types of approaches.

**Table 2.** Types of middleware and their usage

| Type of approach | Networks based on | | |
|---|---|---|---|
| | Sensor | Multimedia | Computation |
| Events | X | | |
| Remote Procedure Call | | X | X |
| Object Request Broker | | X | X |
| Message-oriented | | | X |
| Databases | X | X | |
| Mobile Agents | X | X | X |

It can be noticed that the only approach for middleware that covers all the angles is mobile agents. Among all the agent-oriented middlewares in use today the most widespread is JADE (Java Agent DEvelopment framework). JADE is a completely distributed middleware system [28] with a flexible infrastructure allowing easy extension with add-on modules. The framework facilitates the development of complete agent-based applications by means of a run-time environment implementing the life-cycle support features required by agents, the core logic of agents themselves, and a rich suite of graphical tools. JADE is consistent with the FIPA specifications.

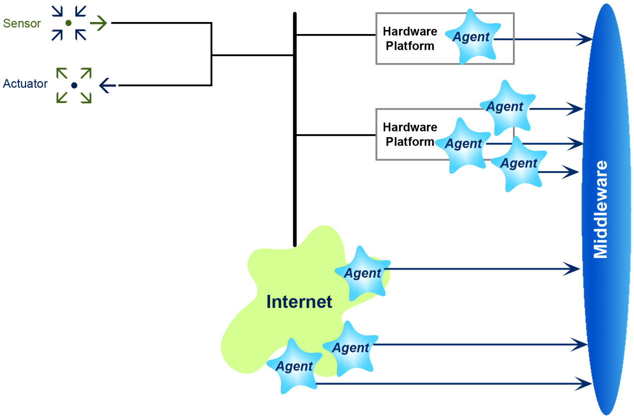


**Fig. 2.** The intelligent grid with intelligent agents

## 2.2 Intelligent Agents

Using intelligent agents allows us to look at the controller/DSP/PC as an agent, with all the implications that the artificial intelligence includes. The agents have several important features like autonomy, proactivity and an ability to communicate. This allows them to execute complex, and often long-term, tasks and to initiate a task even without an explicit stimulus from a user. Thus, the architecture of the overall grid as described in figure 1 on which the concept of intelligent agents is imposed is described in figure 2. Using this agent concept, at the device level we are left only with the sensors and actuators, at the connection level is present the middleware; while at the collection level we have the agents. The agents may run individually on a hardware platform, or may be more than one on a hardware platform.

The interesting thing is that, raising the abstraction level, the physical aspect of the platform is not relevant. Our sole interest is in the agents, and the way they interact and solve the problems. The hardware platform on which an agent is located can be changed at any time, due to agent mobility.

## 3 Redundancy Issues

### 3.1 Redundancy at Sensing Areas Level

While the application potential of networks containing sensors is limitless, the construction of such networks is particularly challenging. One of the main challenges is to maintain long network lifetime as well as sufficient sensing area. First of all, the high density of sensors wastes a lot of energy. A broadly-used approach for reducing energy consumption in sensor networks is to turn off redundant sensors by scheduling sensor nodes to work alternatively based on heuristic schemes [31]. This action might generate blind points and consequently, reduces the network's coverage range.

### 3.2 Redundancy at Link Level

As described in detail in literature, the failure of the transmission medium can not be separated from the failure of the component[32]. But if the same component is redundant linked to several other components we can overcome this particular difficulty. The cost, implied by the extra wires can be justified in certain applications, in which the failure of the transmission medium has a much higher probability than the failure of the component itself.

## 4 Implementation

The hypotesis in which the redundancy at link level was implemented are that the controlers and sensors are in a well known physical area. The sensors in the same area are connected to a controller, but sensor areas overllap in certain degree. An example is shown in figure 3.

This architecture leaves the marginal sensors with less coverage than the middle ones. A possible solution would be the ring connection of the sensors, but that is only possible in specific situations. Anyhow we shall concentrate on the middle part, in which the sensor redundancy/coverage is maximum.

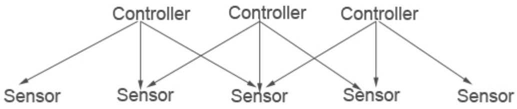


**Fig. 3.** Redundancy at link level

As previously justified, we used JADE. On every controller an agent requests, from time to time, all the other agent's read sensor values. The agents send those values, and the intiator uses all the values in order to set a variable[32]. The transmision wires between the sensors and controllers were supposed to be fail-stop, that is, if an error occurs, the line sends no value. The errors were injected in such manner as to affect as many sensor lines possible. The maximum number of errors in which at least one value of the sensor is recived by any controller is: maximum errors=full covered sensors*(coverage-1).

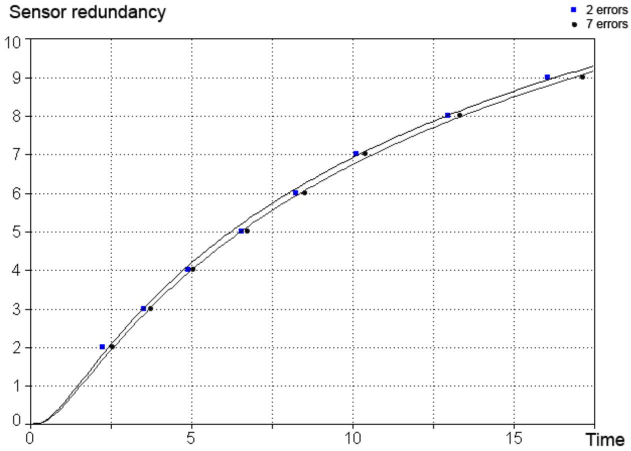


**Fig. 4.** Sensor coverage variable, number of controllers fixed

There are three independent variables implied by this architecture: number of controllers, number of errors and sensor redundancy/coverage The performance metric used to evaluate this variables was execution time. Figure 4, depicting the variation of time with the variation of sensor redundancy while maintaining the number of controllers invariable, shows that the number of injected errors has little effect on the overall execution time. In figure 5 we show the results obtained by maintaining a fixed number of errors (1), while varying the total number of controllers. The effects of enlarging sensor redundancy are obvious, the time increasing with the growth of sensor coverage. It should be mentioned here that the function best describing all this variations is: y=a+bx$^2$.

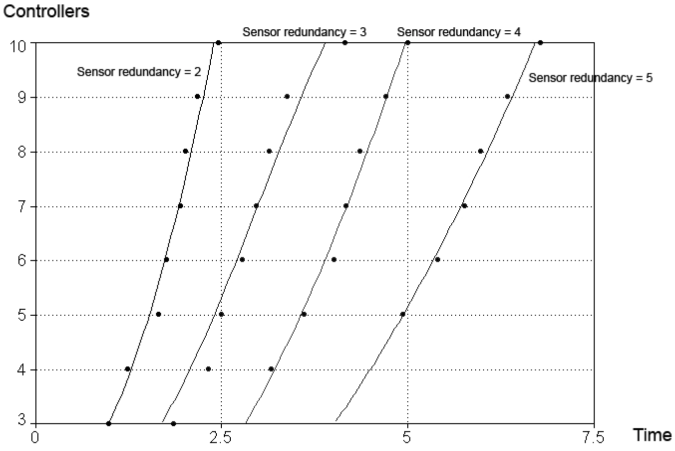


**Fig. 5.** Number of controllers variable, number of errors fixed

An interesting result is shown in figure 6, presenting the time, when variating the total number of controllers and the total number of errors. The first three functions, have the same allure, having a tendency of limiting in the upper section. If the last function, ploted for the situation in which 3 lines are dead, would be charted for a larger number of controllers, the overall outline would be the same asthe others.

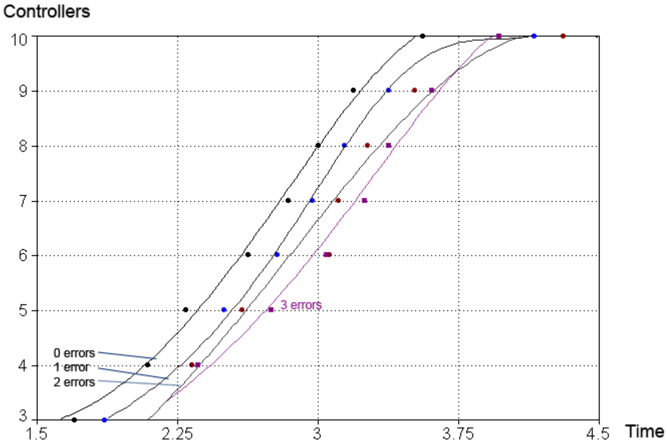


**Fig. 6.** Number of controllers variable, number of errors fixed

## 5 Conclusions

This paper offers a new architecture, of an intelligent grid, based on intelligent agents. The reliability issue when using redundancy is studied. In this regard, the proposed architecture proved to be with a high degree of fault tolerance. This issue was established by error injection in the system. The measurements obtained substantiate this conclusion. Future research will focus on studying the nature of an error and of proposing a method for identifying and correcting the errors.

## References

1. Bell, G.: Bell's Law for the Birth and Death of Computer Classes. Communications of the ACM 51(1), 86–94 (2008)
2. Chong, C.Y., Kumar, S.P.: Sensor Networks: Evolution, Opportunities,and Challenges. Proceedings of the IEEE 91(8) (August 2003)
3. Foster, I., Kesselman, I.: The Grid: Blueprint for a Future Computing Infrastructure. Morgan Kaufmann Publishers, San Francisco (1998)
4. Foster, I.: What is the Grid? A Three Point Checklist. Argonne National Laboratory & University of Chicago (July 2002)
5. Chakrabarti, A.: Grid Computing Security. Springer, Heidelberg (2007)
6. Tham, C.K., Buyya, R.: SensorGrid: Integrating Sensor Networks and Grid Computing. CSI Communications, 24–29 (July 2005)
7. Basu, S., Adhikari, S., Kumar, R., Yan, Y., Hochmuth, R., Blaho, B.: MMGrid: Distributed Resource Management Infrastructure for Multimedia Applications. In: 17th International Parallel and Distributed Processing Symposium, Nice, France (April 2003)
8. Weber, W., Rabaey, J.M., Aerts, E.: Ambient Intelligence. Springer, Heidelberg (2005)
9. Boekhorst, F.: Ambient Intelligence, the Next Paradigm for Consumer Electronics: How will it Affect Silicon? In: ISSCC 2002 IEEE International Solid-State Circuits Conference, vol. 1, pp. 28–31 (2002)
10. Marculescu, R., Rabaey, J.M., Sangiovanni-Vincentelli, A.: Is:Network the Next:Big Idea in Design? In: DATE 2006 Design, Automation and Test in Europe Proceedings, March 2006, vol. 1, pp. 1–3 (2006)
11. Rabaey, J.M.: Wireless Sensor and Consumer Multimedia Networks – A Story of Converging Trajectories? CCNC, Las Vegas (January 2005) (presentation)
12. Weber, W., Braun, C., Dienstuhl, J., Glaser, R., et al.: Disappearing electronics and the return of the physical world. In: 2005 IEEE VLSI-TSA International Symposium on VLSI Technology, April 2005, vol. (25-27), pp. 45–48 (2005)
13. Rabaey, J.M., Burghardt, F., Steingart, D., Seeman, M., Wright, P.: Energy Harvesting – A Systems Perspective. In: Electron Devices Meeting, 2007, IEDM 2007 IEEE International, December 2007, pp. 363–366 (2007)
14. Swanson, R.: Future Developments in Silicon Solar Cells. In: Electron Devices Meeting, 2007, IEDM 2007 IEEE International, December 2007, pp. 359–362 (2007)
15. Venkatasubramanian, R., Watkins, C., Stokes, D., Posthill, J., Caylor, C.: Energy Harvesting for Electronics with Thermoelectric Devices using Nanoscale Materials. In: Electron Devices Meeting, 2007, IEDM 2007 IEEE International, December 2007, pp. 367–370 (2007)
16. Yeatman, E.M., Mitcheson, P.D., Holmes, A.S.: Micro-Engineered Devices for Motion Energy Harvesting. In: Electron Devices Meeting, 2007, IEDM 2007 IEEE International, December 2007, pp. 375–378 (2007)
17. Wang, Z.L.: From Nanogenerators to Nano-Piezotronics. In: Electron Devices Meeting, 2007, IEDM 2007 IEEE International, December 2007, pp. 371–374 (2007)
18. Halla, B.: How the PC Will Disappear. IEEE Computer 31(12), 134–136 (1998)

19. Gates, B.: The Disappearing Computer (2003), http://www.microsoft.com/presspass
20. Roosta, T., Shieh, S., Sastry, S.: Taxonomy of Security Attacks on Sensor Networks. In: First IEEE International Conference on System Integration and Reliability Improvements, Hanoi, Vietnam (December 2006)
21. Chan, H., Perrig, A.: Security and Privacy in Sensor Networks. IEEE Computer 36(10), 103–105 (2003)
22. Romer, K.: Programming Paradigms and Middleware for Sensor Networks. GI/ITG Fachgespräch Sensornetze, Karlsruhe. (February 2004)
23. Bonnet, P., Gehrke, J.E., Seshadri, P.: Querying the PhysicalWorld. IEEE Personal Communications 7(5), 10–15 (2000)
24. Shen, C.C., Srisathapornphat, C., Jaikaeo, C.: Sensor Information Networking Architecture and Applications. IEEE Personal Communications 8(4), 52–59 (2001)
25. Madden, S.R., Franklin, M.J., Hellerstein, J.M., Hong, W.: TAG: a Tiny Aggregation Service for Ad-Hoc Sensor Networks. In: OSDI 2002, Boston, USA (December 2002)
26. Li, S., Son, S.H., Stankovic, J.A.: Event Detection Services Using Data Service Middleware in Distributed Sensor Networks. In: Zhao, F., Guibas, L.J. (eds.) IPSN 2003. LNCS, vol. 2634, pp. 502–517. Springer, Heidelberg (2003)
27. Levis, P., Culler, D.: Mate: A Tiny Virtual Machine for Sensor Networks. In: ASPLOS X, San Jose, USA, October (2002)
28. Boulis, A., Han, C.C., Srivastava, M.B.: Design and Implementation of a Framework for Programmable and Efficient Sensor Networks. In: MobiSys 2003, San Franscisco (May 2003)
29. Bellifemine, F., Caire, G., Greenwood, D.: Developing multi-agent systems with JADE. John Wiley & Sons Ltd, Chichester (2007)
30. Shneidman, J., Pietzuch, P., Ledlie, J., Roussopoulos, M., et al.: Hourglass: An Infrastructure for Connecting Sensor Networks and Applications. Harvard University, Harvard Technical Report TR2104 (2004), http://hourglass.eecs.harvard.edu
31. Avizienis, A., Laprie, J., Randell, B., Landwehr, C.: Basic Concepts and Taxonomy of Dependable and Secure Computing. IEEE Transactions on Dependable and Secure Computing 1(1), 11–33 (2004)
32. Gao, Y., Wu, K., Li, F.: Analysis on the Redundancy of Wireless Sensor Networks. In: WSNA 2003, San Diego, California, USA, September 19 (2003)
33. Pease, M., Shostak, R., Lamport, L.: Reaching Agreement in the Presence of Faults. Journal of the Association for Computing Machinery 27(2)

# A Uniform Parallel Optimization Method for Knowledge Discovery Grid[*]

Kun Gao

Computer Science and Information Technology College, Zhejiang Wanli University
No. 8 South Qianhu Road, China
gaoyibo@gmail.com

**Abstract.** Grid is a new solution to computationally and data intensive computing problems. Since the distributed knowledge discovery process is both data and computational intensive, the Grid is a natural platform for deploying a high performance data mining service. In order to improve the performance of data mining applications, an effective method is task parallelization. Existing mechanisms of data mining parallelization are based on NOW or SMP, it is necessary to develop new parallel mechanism for grid feature. In this paper, we present a framework for high performance DDM applications in Computational Grid environments called Data Mining Grid, with the function for decomposing data mining application into subtasks and then combine those subtasks to form directed acyclic graph. This kind of parallel mechanism decomposes application according to the actual computation power of each node in dynamic Grid environment.

**Keywords:** Grid Computing, Distributed Computing, Data Mining, Knowledge Discovery in Database, Performance Optimization, Parallelization.

## 1 Introduction

Knowledge Grid is software architecture for geographically distributed PDKD (Parallel and Distributed Knowledge Discovery) systems [1]. This architecture is built on top of a computational Grid and Data Grid that provides dependable, consistent, and pervasive access to high-end computational resources[2] [3]. The Knowledge Grid uses the basic Grid services and defines a set of additional layers to implement the services of distributed knowledge discovery on world wide connected computers where each node can be a sequential or a parallel machine.

The Knowledge Grid provides a specialized broker of Grid resources for Parallel and Distributed Knowledge Discovery (PDKD) computations: given a user's request for performing a DM analysis, the broker takes allocation and scheduling decisions, and builds the execution plan, establishing the sequence of actions that have to be performed in order to prepare execution, actually execute the task, and return the results to the user. The execution plan has to satisfy given requirements and constraints-available computing power. Once the execution plan is built, it is passed to the Grid Resource Management service for execution. Clearly, many different

[*] Programs Supported by Ningbo Natural Science Foundation (2008A610028).

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 306–312, 2008.

execution plans can be devised, and the Resource Allocation and Execution Management services have to choose the one which minimizes response time [4, 5]. In order to obtain the minimal response time, an efficient approach for data mining applications is to parallelize the tasks in grid environment.

However, many data mining applications have proved difficult to parallelize, because various pruning mechanisms are used extensively in data mining applications and a powerful pruning mechanism leads to a highly variable search process that conflicts with a uniform workload requirement for good performance [6]. Fortunately, the computation is coarse grain parallel, i.e., it can be parallelized into large, seldom interacting tasks. Coarse grain parallel computations are suitable computations for Grid [7].

In this paper, we propose a new parallel framework on grid for three classes of data mining problems, i.e. association rule mining, classification rule mining, and pattern discovery. According to each existing computing power of grid, the broker can decompose a data mining task to resources available.

The rest of this paper is organised as follows: in section 2, we present the framework that parallelize data mining tasks in Grid environment. In section 3, we present how to optimize the parallel framework. Finally section 6 concludes this paper.

## 2 Related Work

In [8], it is from San Diego Supercomputing Centre which development a middleware to store and access datasets over networks. It is the category of data replication mechanism in fact, for example [9] [10], because it does not handle application implementing in real time. In [11] [12], they are grid computing problem solving environment constructed using MPI and CORBA but is limited to that domain.

In [13], it uses a central server to receive requests and dispatch tasks based on system real time parameters. The main shortcoming of this system is the lack of dynamics. In [14] [15], the systems specialize in parameter-sweep computation, especially supporting dynamic parameters, i.e. parameters whose values are determined at runtime. However, the systems aim at optimizing user parameters and budget for computational tasks only. It has no capability to access remote dataset and optimize the data transfer.

Like [15], [16] provide deployment of parameter-sweep applications on grid. The system emphasizes on data-reuse. The system can appraise the data file that all tasks need, duplicate the data from user node to computation node. When a lot of tasks are assigned to the same resources, it has a try to reuse the data duplicated to make data transmission reduce to minimally. However, the system doesn't support the multiple repositories of data; this method is not applicable to grid.

In this paper, we present a framework for high performance DDM applications in computational grid environments called DMGrid, which is based on grid mechanisms and implemented on top of the Globus 4.0 toolkit. It integrates Grid services by supporting distributed data mining, task scheduling and resource management services that will enlarge the application scenario and the community of Grid computing users.

## 3 The Rationale for Design and Development of the Data Mining Grid

The rationale of design and development the grid enabled data mining system is as follows:

- DMGrid adopts the standard, common and open grid service mode, follows OGSA norm, and offers unified support to the data mining applications.
- Based on Globus Toolkit and according to the existing networks system structure, DMGrid use the grid service to realize communication, operates each other and resource management.
- DMGrid is open, supports various data mining tools and algorithms, the extensibility is good.
- DMGrid is able to realize the improvement of performance by increasing network node, high performance computing node and cluster, the scalability is strong.
- DMGrid can deal with distributed huge volumes of high dimensional dataset, support heterogeneity data source.
- The main purpose to design and develop the DMGrid system is to improve the performance.
- Users carry out the data mining tasks in a transparent way; the concrete system structure, operation and characteristic in the grid environment is to be hidden.
- In the field of data mining, the security of the data and personal secrets are a sensitive topic. Data Ming Grid supports the choice of place that the data mining execute.

## 4 The Structure of the Data Mining Grid

Fig. 1 describes the distributed system framework that we designed and developed. It is mainly made up by following components:

- DMGrid Client Node: In consideration of ease of use, the system adopts Browser/Server mode. Grid client exchanges information with Grid portal through Internet Explorer browser. Users submit the requirement of data mining and receive the final result at Grid client.
- DMGrid Portal Node: It provides a single access way to distributed data mining application based grid. Users can make use of the whole grid resource transparently through the grid portal. This component is responsible for translating users' demand into the RSL language (Resource Specification Language) that can be recognized by grid, is used for grid resource discovery and grid resource allocation management. The final result is returned to grid portal first, and then returned to users by the portal.
- DMGrid Resource Broker Node and DMGrid Tasks Allocation Broker Node: user’s data mining requirement has driven grid resource discovery. According to users' demand condition, DMGrid resource broker looks for the resources which

meet the condition in a large number of grid resources, including algorithms, computing capability and data resource. It is an important job that finds appropriate resource [11] [12]. As to any application based on grid, it is first to find appropriate resource, then allocate tasks and management them. It can be predicted that there may be many nodes which meet a condition. Resource broker is used for finding available resource among MDS (Meta Directory Service); mapping between data resource and computing resource, i.e., the task allocation broker is responsible for dispatching a certain task on a certain node.

- Grid Node: The Grid nodes are made up of personal computer, high performance computer and cluster. Each node is installed GLOBUS, as grid middleware. They are the data carrier and the computation implementation entity.

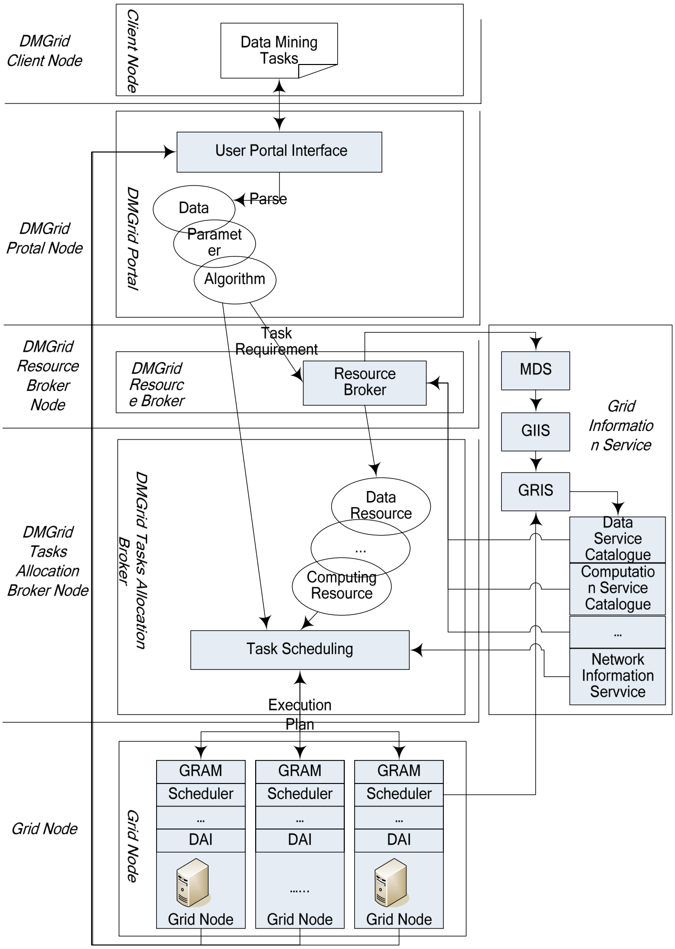


**Fig. 1.** The framework of distributed data mining on grid

# 5 Parallelizing Data Mining Tasks in Grid Environment

DMGrid services can be used to build parallel applications solving environment. It can decompose a complex data mining application to some coarse grain subtasks and describe those subtasks as a Directed Acyclic Graph (DAG). In this DAG, the nodes are data mining algorithms and data sets, and the links are relationship of data sets. In this section, we present how to map data mining application to DAG.

## 5.1 Uniform Description for Data Mining Applications

We study main classes of data mining applications, that is, association rule mining, classification rule mining, and pattern discovery and so on. We note the resemblance existing in the data mining computation models. Table 1 is a comparison of specifications of these three classes of applications.

A certain data mining application is computing on a pattern. Not only are all tasks of any one application of the same kind, but tasks of different applications are actually very similar. They all take a pattern and a subset of the database and count the number of records in the subset that match the pattern. In the classification rule mining case, counts of matched records are divided into c baskets, where c is the number of distinct classes.

**Table 1.** A comparison of specifications of three classes of data mining applications classes

| | Pattern Discovery | Assoc. Rule Mining | Class. Rule Mining |
|---|---|---|---|
| Data-base | Sequences | Transaction records | Database relation |
| Pattern | Partial sequence | Itemset | Attribute-value condition |
| Good pattern | $\text{occurrence}_{\text{pattern}}>$ min_occurrence | $\text{support}_{\text{pattern}}>$ min_support | Info_$\text{gain}_{\text{attribute}}=$ $\text{Max}_{\text{sibling}}$_attributes (info_$\text{gain}_{\text{attribute}}$) |
| Task | Countingoccurrence of pattern in subset of database | Counting support of itemset over subset of database | Building histogram of pattern on classes over subset of database |

The similarities among the specifications of these applications are obvious, which inspired us to study the similarities among their computation models. They usually follow a generate-and-test paradigm-generate a candidate pattern, then test whether it is any good. Furthermore, there is some interdependence among the patterns that gives rise to pruning, i.e., if a pattern occurs too rarely, then so will any super-pattern. These interdependences entail a lattice of patterns, which can be used to guide the computation.

In fact, this concept of pattern lattice can apply to many data mining application that follows this generate-and-test criterion. We call this application class as data mining based on lattice theory.

We propose to use a directed acyclic graph structure called Data Mining DAG, DM-DAG for short, to characterize data mining applications based on lattice theory.

The DM-DAG build for data mining applications has as many vertices as the number of all possible patterns. Each vertex is labeled with a pattern and no two vertices

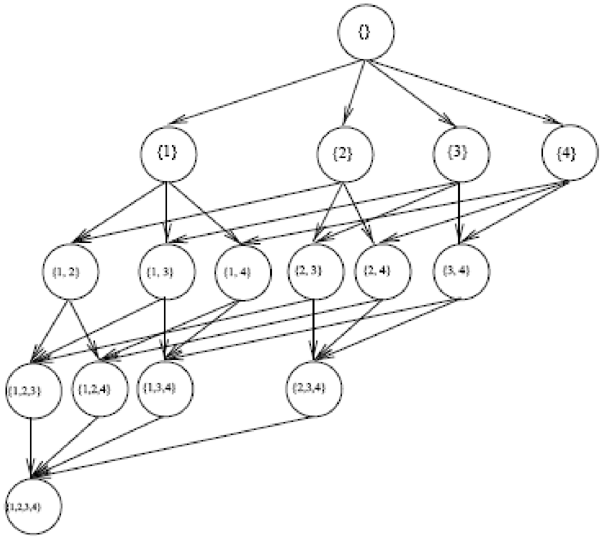


**Fig. 2.** A DM-DAG for an association rule mining application

are labeled with the same pattern. Hence there is a one-to-one relation between the set of vertices of the DM-DAG and the set of all possible patterns. Therefore, we refer to a vertex and the pattern it is labeled with interchangeably. Fig. 2 shows a DM-DAG.

## 6 Conclusions

Data mining is a nontrivial process of computation. An efficient data mining application should overcome many factors affecting its performance. With the emergence of globalization grid computing platform, it causes large-scale resource sharing and cooperation becomes possibly. This brings the new vitality for the development of distributed data mining. The grid can provide the high expensive computation resources which the data mining needs. At the same time, the grid environment is consistent with the inherent property of distribution and heterogeneity of data. It is a new trend to merge grid and data mining to meet demands of applications. In this process, optimizing the strategy of tasks allocation is extremely important feasible way to improve the performance of distributed data mining application.

## References

1. Talia, D., Cannataro, M.: Knowledge grid: An architecture for distributed knowledge discovery. Communications of the ACM (2002)
2. Foster, I., Kasselman, C.: The Grid: blueprint for a future infrastructure. Morgan Kaufman, San Francisco (1999)
3. Chervenak, A., Foster, I., Kesselman, C., Salisbury, C., Tuecke, S.: The Data Grid: towards an architecture for the distributed management and analysis of large scientific datasets. J. of Network and Comp. Appl. (23), 187–200 (2001)

4. Gao, K., Chen, K., Liu, M., Chen, J.: Rough Set Based Data Mining Tasks Scheduling on Knowledge Grid. In: Szczepaniak, P.S., Kacprzyk, J., Niewiadomski, A. (eds.) AWIC 2005. LNCS (LNAI), vol. 3528, pp. 150–155. Springer, Heidelberg (2005)
5. Gao, K., Ji, Y., Liu, M., Chen, J.: Rough Set Based Computation Times Estimation on Knowledge Grid. In: Sloot, P.M.A., Hoekstra, A.G., Priol, T., Reinefeld, A., Bubak, M. (eds.) EGC 2005. LNCS, vol. 3470, pp. 557–566. Springer, Heidelberg (2005)
6. Talia, D.: Parallelism in Knowledge Discovery Techniques. In: Fagerholm, J., Haataja, J., Järvinen, J., Lyly, M., Råback, P., Savolainen, V. (eds.) PARA 2002. LNCS, vol. 2367, pp. 127–136. Springer, Heidelberg (2002)
7. Cannataro, M., Srimani, P.K., Talia, D.: Parallel Data Intensive Computing in Scientific and Commercial Applications. Parallel Computing 28(5), 673–704 (2002)
8. Samar, A., Stockinger, H.: Grid Data Management Pilot (GDMP): A Tool for Wide Area Replication. In: IASTED International Conference on Applied Informatics (AI 2001), Innsbruck, Austria (February 2001)
9. Baru, C., Moore, R., Rajasekar, A., Wan, M.: The SDSC Storage Resource Broker. In: CASCON 1998 Conference, Toronto, Canada (1998)
10. Chervenak, A., Deelman, E., Foster, I., Guy, L., Hoschek, W., Iamnitchi, A., Kesselman, C., Kunst, P., Ripeanu, M., Schwartzkopf, B., Stockinger, H., Stockinger, K., Tierney, B.: Giggle: A Framework for Constructing Scalable Replica Location Services. In: Proceedings of Supercomputing 2002 (SC 2002) (November 2002)
11. Allen, G., Benger, W., Goodale, T., Hege, H., Lanfermann, G., Merzky, A., Radke, T., Seidel, E., Shalf, J.: The Cactus Code: A Problem Solving Environment for the Grid. In: Proceedings of the Ninth International Symposium on High Performance Distributed Computing (HPDC), Pittsburgh, USA. IEEE Press, Los Alamitos
12. Marzullo, K., Ogg, M., Ricciardi, A., Amoroso, A., Calkins, F., Rothfus, E.: NILE: Wide-Area Computing for High Energy Physics. In: Proceedings of 7th ACM SIGOPS European Workshop, Connemara, Ireland, September 2-4. ACM Press, New York (1996)
13. Hoschek, W., Jaen-Martinez, J., Samar, A., Stockinger, H., Stockinger, K.: Data Management in an International Data Grid Project. In: Proceedings of the 1st International Workshop on Grid Computing (Grid 2000, Bangalore, India). Springer, Berlin, Germany (2000)
14. Abramson, D., Giddy, J., Kotler, L.: High Performance Parametric Modeling with Nimrod/G: Killer Application for the Global Grid? In: Proceedings of the International Parallel and Distributed Processing Symposium (IPDPS 2000), Cancun, Mexico, May 1-5. IEEE CS Press, USA (2000)
15. Buyya, R., Abramson, D., Giddy, J.: An Economy Driven Resource Management Architecture for Global Computational Power Grids. In: Proceedings of the 2000 International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA 2000), Las Vegas, USA, June 26-29. CSREA Press, USA (2000)
16. Casanova, H., Obertelli, G., Berman, F., Wolski, R.: The AppLeS Parameter Sweep Template: User-Level Middleware for the Grid. In: Proceedings of the IEEE SC 2000: International Conference Networking and Computing, November 2000, Dallas, Texas. IEEE CS Press, USA (2000)

# E-Service Delivery Pattern for Knowledge Discovery Grid*

Meiqun Liu

Culture and Communication, Zhejiang Wanli University
No. 8 South Qianhu Road, Ningbo 315100, China

**Abstract.** The emergence of ASP hosting Grid based data mining services is being seen as a novel and feasible solution for organizations that value their knowledge resources but are constrained by the high cost of knowledge discover tools. In this paper, we present a novel E-service delivery pattern for the Grid based data mining services. This pattern has several strong suits than the present methods for delivering data mining services such as supporting clients with many kinds of choices, especially the choice of the ASP. We have developed tools to support the E-service delivery pattern. The detailed information is including: the preferences of tasks, the functionality of ASP and the information of data resource and computation resources to run the application. We have realized the map function to match the requirement of client, the capability of resource and ASP.

**Keywords:** Grid Computing, Distributed Computing, Data Mining, Knowledge Discovery in Database, Grid service.

## 1 Introduction

ASP, Application Service Provider, is more and more focused on the knowledge as a key to support DSS (decision support system) service for the clients. Data mining is a main method for extracting knowledge. With the development of Grid computing technology[1,2,3,4], data mining service taking Grid as platform has appeared. This kind of data mining service based Grid delivery has appeared as an is emerging as an magnetic alternative for many institution, which are the most restrained by the high expense of knowledge discovery tools. The greatest advantage of this kind of service is for using the software to pay, and does not need to undertake the cost for buying, installing and training. ASP emphasizes the survival ability of hire tools [5]. The client should not buy a high cost software tool and set up, but only hire the usage of some software tools of ASP. DM has several features which let it to conform to the ASP pattern. The appropriate characteristics for providing knowledge discovery are as bellow:

- Commercial application can be various demands from client characteristic to market decision. Such variety demands knowledge discovery systems that could support more solution for meeting the requirements. Data mining systems have been developing from only one algorithm to complex system including many

* Programs Supported by Ningbo Natural Science Foundation (2008A610028).

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 313–317, 2008.

algorithms, lots of customers, different data formats and distributed data set. This trend of development is different from others decision-making system which can provide all commercial decision. ASP can lighten this question through integrating many kinds of data mining systems of totally different needs satisfied with customers.

- Applications like E-business support a framework to integrate many organizations requirement to support contend in area domain, which were former wide organization and multi-country. These organizations want to obtain income by using intelligence tools from the information resources and thereby gain a competitive edge. The small-scale company is not fit for the expensive expenses of this kind of software. ASP is a kind of better solution for this case.
- An organization obtains the income is relatively long by using a completed suit of data mining tools in its inside. Because this needs a long time to learn curve concerned the usage of knowledge discovery tools. If one requires benefits at once, the usage of ASP is the best way.

## 2 Related Work

### 2.1 DigiMine

DigiMine [6] is setting new standards for the delivery of powerful data mining and analytics. DigiMine managed data mining solutions transform raw data into actionable business intelligence for more profitable marketing campaigns, sales interactions and customer relationships. The company's solutions are powered by a managed data warehouse and delivered via the Internet, providing fast deployment, a low total cost of ownership and outstanding return on investment. DigiMine uncovers valuable business intelligence and enables companies to take immediate action by delivering advanced analytics and personalization tools.

DigiMine Inc. is tackling an underserved area of data mining: wireless. The company, founded by three Microsoft defectors, has a new service called Wireless Business Intelligence that analyzes log-file data for wireless content providers, service providers, and carriers.

DigiMine provide daily reports to help clients fine-tune their efforts in wireless content, commerce, and products. For example, clients could have the data analyzed to determine what type of content is best suited for certain wireless-user demographics, or why wireless shoppers abandon their shopping carts at particular points.

The wireless industry, which suffers from severe customer churn, could certainly use some assistance identifying relevant content for users. It's not that large wireless providers are lacking sophisticated data warehouses. The problem is that many of the data warehouses weren't designed to let those providers act quickly on information, with reports generated in batches or only periodically.

### 2.2 Kinetic

Kinetic[7] Services has been providing distinct value through full lifecycle data warehousing and business intelligence solutions. Through the unique skill sets, a consistent, proven approach and a set of rigorous controls, Kinetic have built a solid reputation and a foundation for success from which it continue to build.

Kinetic Services offers the following solutions:

- Strategy and Audit Services
- Full Data Warehouse and BI Life-Cycle Design and Build
- Center of Excellence Creation and Support
- Remote Hosting, Management and Support

Kinetic is designed with an impressive suite of reusable components. Kinetic Services provide knowledge transfer, training and documentation to empower them as systems evolve.

In order to provide an alternative and make products available to more of the market, Kinetic Networks decided to open-source its production software of Kinetic Networks Extract Transform and Load tool (KETL). This ETL tool allows companies to manage the complex manipulation of data while leveraging the affordability of an open source data integration platform. KETL's flexible and production-proven ETL engine provides the foundation for successful data warehousing or data integration efforts. Its comprehensive feature set is furthered enhanced by open source participation to provide superior capabilities at a lower entry cost.

### 2.3 Webminer

WebMiner ASP [8] is a subscriber-based data mining tool for e-commerce Web sites. By evaluating both customer databases and clickstream data, rules are generated that are used to segment users allowing customization for future Web site visitors.

All the data in Webminer is freely available from SGD or published works cited in Webminer. Webminer makes it easy to perform sophisticated searches by organizing some of the vast wealth of published genomic data under one roof.

Webminer includes a variety of information about every ORF in the yeast genome: its promoter, the protein it encodes, and how its mRNA levels change under many different conditions.

## 3 Concerned Concept

A variety of technologies that support Web services is starting to appear. The leading candidates include SOAP, SOAP Messages with Attachments, WSDL, UDDI, and ebXML. Currently, these technologies are not available as supported products, but they are maturing quickly.

### 3.1 SOAP

SOAP (Simple Object Access Protocol) is technology developed by DevelopMentor, IBM, Lotus, Microsoft, and UserLand. SOAP provides an extensible XML messaging protocol, and it provides an RPC programming model application layer. At this time, more than 30 SOAP implementations are available. The two most popular implementations are an open source Java implementation from the Apache Software Foundation and a Microsoft implementation within the .Net SDK.

### 3.2 SOAP Messages with Attachments

SOAP Messages with Attachments, developed by Hewlett Packard and Microsoft, defines a binding for a SOAP 1.1 message to be carried within a MIME multipart/related message, enabling a SOAP message to carry multiple XML documents and non-XML attachments. Microsoft BizTalk Server uses SOAP Messages with Attachments, so this technology is included in a supported product.

### 3.3 WSDL

WSDL is an XML format for describing network services as a set of endpoints operating on messages containing either document-oriented or procedure-oriented information. The operations and messages are described abstractly, and then bound to a concrete network protocol and message format to define an endpoint. Related concrete endpoints are combined into abstract endpoints (services). WSDL is extensible to allow description of endpoints and their messages regardless of what message formats or network protocols are used to communicate, however, the only bindings described in this document describe how to use WSDL in conjunction with SOAP 1.1, HTTP GET/POST, and MIME.

### 3.4 UDDI

The Universal Description, Discovery, and Integration (UDDI) specification defines a SOAP-based Web service for locating Web services and programmable resources on a network. UDDI provides a foundation for developers and administrators to readily share information about internal services across the enterprise and public services on the Internet.

### 3.5 eb-XML

ebXML (Electronic Business XML) is a B2B XML framework being developed by the ebXML Initiative. The ebXML Initiative is a joint project of UN/CEFACT (the United Nations body for Trade Facilitation and Electronic Business) and OASIS (Organization for the Advancement of Structured Information Standards). The ebXML membership includes representatives from more than 2000 business, governments, institutions, standards bodies, and individuals from around the world. ebXML is a complete B2B framework that enables business collaboration through the sharing of Web-based business services.

## 4 E-Service Delivery Pattern for Data Mining

The E-service Delivery Pattern shows in fig.1.

This pattern can effectively get over the shortcoming of existing delivery pattern to match the demand of applications. The procedure of this pattern follows three steps:

1. The specification among clients and ASPs must be exchanged
2. The ASP searches the corresponding nodes for matching the specification of client, computing resource, data set resource and concerned algorithm;
3. The portal selects some nodes that can obtain the best benefits to execute the application.

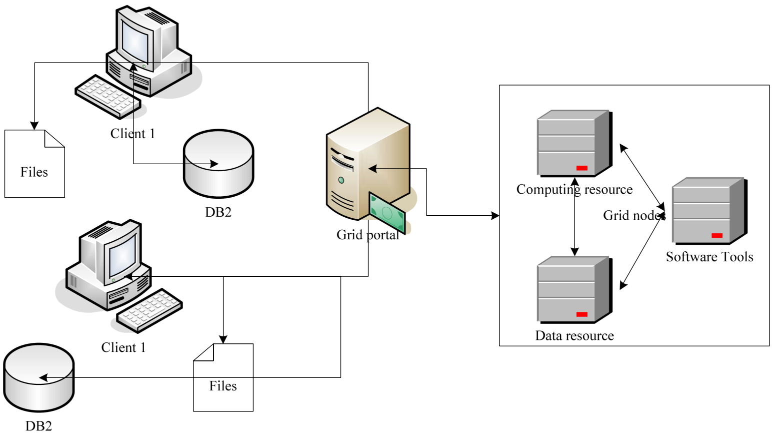


**Fig. 1.** E-service Delivery Pattern

## 5 Conclusion and Future Works

The emergence of Grid based ASP service is to be a feasible solution for meeting the applications intelligence demands of knowledge discovery. In this paper we have presented a novel ASP pattern to overcome the shortcoming of existing data mining application. We have also presented the benefits of this solution pattern. It can support the preferences of clients, ASP and resources.

At present, our works focus on the ranking schemes for ASP. We believe that ranking, along with negotiation techniques for Grid services are very important research problems in this direction.

## References

1. Foster, I., Kasselman, C.: The Grid: blueprint for a future infrastructure. Morgan Kaufman, San Francisco (1999)
2. Talia, D., Cannataro, M.: Knowledge grid: An architecture for distributed knowledge discovery. Communications of the ACM (2002)
3. Chervenak, A., Foster, I., Kesselman, C., Salisbury, C., Tuecke, S.: The Data Grid: towards an architecture for the distributed management and analysis of large scientific datasets. J. of Network and Comp. Appl. (23), 187–200 (2001)
4. Cannataro, M., Srimani, P.K., Talia, D.: Parallel Data Intensive Computing in Scientific and Commercial Applications. Parallel Computing 28(5), 673–704 (2002)
5. Clark-Dickson, P.: Flag-fall for Application Rental, Systems, August, pp. 23–31 (1999)
6. `http://www.digimine.com`
7. `http://www.kineticnetworks.com/`
8. `http://www.webminer.com/`

# Design and Implementation to Intelligent Examination System for e-Business Application Operation

Xin Jin and YanLin Ma

School of Information, Central University of Finance& Economics
Beijing, 100081, P.R. China

**Abstract.** This paper presents a design and implementation of an intelligent testing system for checking the e-business application operation capable of the examinees. A novel architecture for on-line examination system is proposed. That system adopts the common client-server pattern with two major parts, the client component also called Student site and the application-oriented server component here also called Teacher site. The Student site provides examinee a GUI (Graphics User Interface) embedded a WWW browser for answering the paper. The Teacher site mainly includes paper management subsystem for managing paper database, a building paper subsystem for building a new paper, and a scoring paper subsystem for scoring the paper of the examinees. In this paper, we also analyze the system architecture, some key questions and the corresponding solutions.

**Keywords:** We would like to encourage you to list your keywords in this section.

## 1 Introduction

The use of information technology (IT) has become a primary survival factor for business organizations in a global competitive environment, as the e-business tide is spreading through various domains violently in modern society. Developing e-business is an important factor to accelerate national economy increasing.

In order to adapt and grasp this situation and respond the proposal in the Eleventh Five Planning of Nation “accelerate developing e-business, apply information technology widely and popular information knowledge and craftsmanship widely in the whole society.”. So the education and the training of the e-business knowledge are actively developed in whole society. Under that great background, we develop the Examination System for e-Business Application Operation (ESBAO) which is a part of e-business education and training software system supported by Shanghai Informatization Office, so as to check the user’s ability of the e-business application operation.

Commonly, the traditional e-business education and training adopts two ways: the theoretics teaching in classroom and the computer-based application operation in e-business simulation environment which is like as e-business web site (but not same as).

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 318–323, 2008.

Here, We will introduce examination system that is for checking the examinee's ability of e-business application operation. This system has some characters as follows:

*I. Checking the ability to e-business application operation of the students or the examinees*

The ESBAO system gives marks for the students according to whether their operations which include the key operation steps and the operational results are correct or not under the e-business simulation environment. For example, if the question is buying commodity A by the searching way, you can do the correct answer as following workflow: *start up searching engine -> select commodity A -> start up purchasing engine -> make orders -> finish*. Your answer also can be completed through the other workflow. This system gives marks for you just according to whether starting up searching engineer and successfully purchasing A or not.

*II. Intelligence*

The intelligence of ESBAO is embodied mostly with the intelligence of building examination paper subsystem and scoring examination paper subsystem in Teacher site. For example, the teacher can building the examination paper by himself through GUI, also he can only start up the Building Paper Agent (BPA) which can build examination paper automatically. When the teacher scores the examination papers, he only needs to star up scoring paper engine through GUI, then the system would check and mark the examination papers and record the examinees' scores in the database.

The rest of the paper is organized as follows: In section 2 we introduce the design of the ESBAO system architecture. In section 3 discuss the system design and implementation essential. In section 4, we discuss the key technologies for system implementation. Finally it is the conclusion of this paper.

## 2 System Architecture

We use the XML technology[1], the Component technology[2], the Database Trigger and Stored Procedure technology[3] and other information technology to design and realize the ESBAO. This system mainly comprises Paper Database, Paper Creator, Building Paper Agent, Monitor for monitoring application operations of the examinees, Answer Generator and Scoring Paper Generator. The Fig.1 depicts the system architecture and the workflow.

In Fig.1, the Paper Database is collector of examination questions described by XML document style. The teacher can simply and quickly creates a new XML document paper, only need to start up Paper Creator component engine that can work automatically. The Paper Creator is very expediently for the teacher to parser XML document paper to automatically analyze and deal the paper's structure, attributes, scores, contents etc. When the teacher finishes building paper, the corresponding standard answers are automatically created. These building paper procedures also can work automatically by Building Paper Agent, after the teacher start up the agent.

When the examinees answer the questions of the paper, it is necessary to operating in the e-business simulation environment which is like but not same as the e-business web site. The main difference is that e-business simulation environment sets a monitor program in some main operation steps to capture who and whenever and however to operating. The monitor program is oriented special operation workflow, and it is sequential and discretely in workflow and in work time, so as to the system records the operations which are the some key operation steps or result in database. When scoring the papers, we can use Answer Generator to transfer the recorders of the database to XML formatted answer sheets. In this system, all of documents are XML-based format.

When the teacher starts up the Scoring Paper Generator which can score papers automatically, it can compare examinees' answers with the standard answer to give marks. Because the all document are uniform XML format, it only needs to judge the data identity, then to mark the scores.

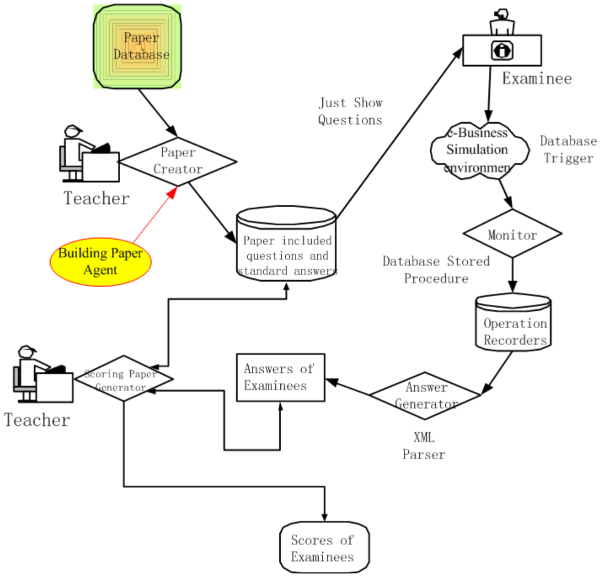


**Fig. 1.** System Architecture and Workflow of the ESBAO

## 3 Discuss the System Design and Implementation Essentials

The ESBAO system can check the examinee's application operation procedures and results under the e-business simulation environment. Its designing goal is to test e-business application operation ability.

It is necessary to solve some important questions as follows:

- capturing and analyzing the operations
- concurrency of many operations
- building paper and scoring paper automatically

### 3.1 Capturing and Analyzing the Operations

The most methods to recording the operation procedures are relative with the structure of the e-business simulation environment. If it is Client/Server structure, the

ESBAO system can set monitor program in client and server together. If it is Browser/Server structure (this system uses this structure in fact), the system can only set monitor program focused in server to search and capture the operations of the client. As the above mentioned, this monitor methods are discretely in the business workflow. So before building monitor program centralized in server, we must firstly be confirmed the monitored objects. That means we need to confirm which operation steps and results need to be monitored.

The simply solution is to add triggers in database to monitor add, delete and update operations. These data changing often be from the activity and the operation in business environment. Furthermore, because the system design is based on server component program, we can add log records in component to capture operations. We can add an examination monitor switcher for all components. When beginning the examination, the system open this switcher, then the operations and activities can request these components to create some fixed format records which can reflect the operation types.

Furthermore, if some contexts of simulation environment has used middleware or middleware platform[4], we can set a series of monitor program to realize the operation capturing and analyzing.

## 3.2 Concurrency of Many Operations

At current, the concurrency is not very difficult to implement contrasted to past time, because the most of commerce software and the system program developing architecture provide responding solution.

But in the ESBAO system, because there are many examinees to take part in examination together in a server e-business simulation environment, so there are some questions to solve as follows. For example:

- When we operate a certain defined resource, if the resource is not database (the database can balance the concurrent operation) but data file, media or print device, Once the resource be locked by one student's operations, the other students can not use that resource to finish the operation. Besides the later operation could overlap the previous operation, so some different meanings are arose.
- In some typical e-business application workflow, some enterprise entity application, especially the application come down to financing, auction, sale or paying etc., often be sensitive to the activity entity. For example, auditing the order, the auditing order can not be looked through by other people. The kind of operations, such as the above operation of auditing order, can not fit as the examination questions.

Besides some e-business application workflows often need at least two aspects to finish. For example, the online auction in which there are some business entity which often affect and restrict each other. So, to check many people cooperation operation is an interesting approach issue.

In this system, we can adopt the following ways to solve the above concurrent questions.

- Adopt memory database and data view

In the common situation, the data of the examinee operation often distribute in temp data cache and view but not operate directly in the center database. Then the data of examinee operation can not be interfered each other.

- Isolate the public data and private data

In some typical workflow, the system can create some backups of used data and resource. Different examinees can use different backups. It is not need to backup the data which is not relative with these operations.

- Using programs to simulate many people cooperation operation

For example, in the contesting auction, the system adds the dummy contest rival to contest with the examinee.

- Data restore

After the examination, in order to keep the stability and standardization, the system provided toolkits which can restore the data and the resource to the primal status.

### 3.3 Manage Papers and Score Papers Automatically

In order to manage the papers, using the paper database is a good solution. In the database, all examination questions have been test to be sure the correctness in the examination procedure. To the examination in simulation environment based on the components and comprised of some work units, the teacher can use the papermaking toolkit to build the questions and do parameter setting in operation units of the integrate workflow to form a series of operation rules. The component and the work units can judge the mark referring those rules, so those methods can make the questions type and formal to be more diversified.

In the ESBAO system, besides the checking paper automatically based on XML documents compared, we add the personality requirement and evaluating guidelines, which make the marking automatically to be more impersonality. In most situations, e-business operations can not record and weight by estimated way. For example, using the searching engine, we can search one commodity through various ways. In that procedure, it is difficult to judge the operation is correct or not. Those lead to the process-typed questions and result-typed questions appeared. The called process-typed question means to emphasize the sequence of a series of operations, but the result-typed questions only emphasize the final result. For those two different questions types, we need use different judging standards to deal. So in the process-typed questions, we can set some monitor points in the operation procedure, also we can divide the integrated workflow to many sub objects. The system not stickle to the real application workflow, but add more require and evaluation guidelines artificially. This mark based on the subdivision workflows is apparently more correct.

## 4 Key Technologies for System Implementation

On the above discussion, we analyzed the system design requirement. Because of the characters of testing contexts and the intelligence requirement, we adopt some technologies to implement the system as follows:

- Component technology

Considered the characters of the operation mode, the design based on components gives priority of business- oriented or procedure- oriented. In order to combine conveniently the script language of simulation environment (in this system, our simulation environment adopted the ASP script), we use the C++ and VB to compile the COM we needs. For example, in order to record the examinees' answers, we need to set monitor program to monitor that operation procedure. In ESBAO system , we use the component technology to enclose most functions to strengthen the software reused.

- XML technology

In this system, we use XML-based documents to store the data of the papers, and use XML technology, such as XML DOM ,XML SAX and ADO, to realize the dynamically mapping from the XML documents to database recorders, transferring information and maintaining information etc.

- Database technology

One hand the database is the carrier of the data storage, the other hand the system can use the database triggers and/or the component events to stimulate the database stored procedures to record the operations of the examinees.

- Agent technology[5]

In this system, the Building Paper Agent is an automatic component in fact, in which we add the message mechanism to realize the autonomy of the agent. We use C++ to implement the agent.

## 5 Conclusion

In this paper, we discuss an operation-oriented intelligence examination system, which is for e-business application operation examination, and present a novel system architecture for the Examination System for e-Business Application Operation (ESBAO). The ESBAO system mainly include an answering paper subsystem for examinees in the Student site, paper management subsystem in Teacher site for managing paper database, a building paper subsystem in Teacher site for building a new paper, and a scoring paper subsystem in Teacher site for scoring the paper of the examinees. In this paper, we detailedly discuss the system structure, some key questions and the corresponding solution.

## Acknowledgement

This research was supported by Natural Science Foundation of Beijing under Grant No. 4073041.

## References

1. Blex, udC: XML Technology. Tsinghua University Press, Beijing (2003)
2. Kirtland, M.: Application Program Based on COM. Peking University Press, Beijing (1999)
3. Popa, L., Velegrakis, Y.: Translating Web Data. In: Proceeding of the 28th VLDB Conference, HongKong, China (2002)
4. Lenzerini, M.: Data integration: A theoretical perspective. In: PODS, pp. 233–246 (2002)
5. Shi, Z.: Intelligent Agent and Application. Science Technology Press, Beijing (2001)

# Application Study in Decision Support with Fuzzy Cognitive Map

Yue He

Department of Computer, Shaoxing College of Arts and Sciences
Shaoxing Zhejiang 312000, China

**Abstract.** Fuzzy cognitive map is an approach to knowledge representation and inference. Little research has been done on the goal-oriented analysis with FCM. We propose a methodology based on the use of Fuzzy cognitive map and immune algorithm to find the initial state of system from among a large number of possible states. Finally, an illustrative example is provided, and its results suggest that the method is capable of goal-oriented decision support.

## 1 Introduction

FCM is a soft computing method for simulation and analysis of complex system, which combines the fuzzy logic and theories of neural networks. It offers a more flexible and powerful framework for representing human knowledge [1, 2] and reasoning. FCM have been used for representing knowledge and artificial inference, and have found many applications, for instance, geographic information systems [3, 4], fault detection [5], policy analysis [6], etc. Little research has been done on the goal-oriented analysis with FCM. In this paper, we propose a methodology based on the use of Fuzzy cognitive map (FCM) and immune algorithm for goal-oriented decision support. It aims to provide a means for goal-oriented decision support.

The paper is organized as follows. Section 2 presents the formalization and the inference process of FCM. Section 3 presents a brief overview of the immune algorithm. Section 4 presents how to use FCM for goal-oriented decision support. Section 5 applies the proposed methodology to goal-oriented analysis. Section 6 is the conclusion and suggestions for future works.

## 2 Fuzzy Cognitive Map

### 2.1 Formalization of Fuzzy Cognitive Map

The graphical illustration of FCM is a signed directed graph with feedback, which is consisted of nodes and weighted arcs. Nodes of the graph stand for the concepts that are used to describe the behavior of the system and they are connected by signed and weighted interconnections representing the causal relationships. A FCM is a method to draw a graphical representation of a system, and consists of nodes-concepts, each node-concept represents one of the key-factors of the system, and it is characterized

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 324–331, 2008.

by a value C$\in$ (0,1), and a causal relationship between two concepts is represented as an edge wij. $w_{ij}$ indicates whether the relation between the two concepts is direct or inverse. The direction of causality indicates whether the concept $C_i$ causes the concept $C_j$.

A FCM is a 4-tuple (***V, E, C, f***) where --***V***={$v_1$, $v_2$, ... , $v_n$} is the set of n concepts forming the nodes of a graph.

***E:***($v_i$, $v_j$) $\rightarrow$ $\boldsymbol{w_{ij}}$ is a function $\boldsymbol{w_{ij}}\in$ ***E***, $v_i$, $v_j$ $\in$ ***V***, with $\boldsymbol{w_{ij}}$ denoting a weight of directed edge from $v_i$ to $v_j$. Thus ***E*** ***(V***$\times$***V)=(***$\boldsymbol{w_{ij}}$***)*** is a connection matrix.

***C:*** $v_i$ $\rightarrow$ $\boldsymbol{C_i}$ is a function that at each concept $v_i$ associates the sequence of its activation degrees, such as $\boldsymbol{C_i(t)}$ given its activation degree at the moment *t*. ***C(0)*** indicates the initial vector and specifies initial values of all concept nodes and ***C(t)*** is a state vector at iteration *t*.

***f*** is a transformation function, which includes recurring relationship between ***C(t+1)*** and ***C(t)***.

$$\mathrm{C}_{\mathrm{i}}(\mathrm{t}+1) = \mathrm{f}(\left(\sum_{\substack{i=1\\ j\neq i}}^{n} w_{ij} C_{j}(t)\right)) \tag{1}$$

the transformation function is used to confine the weighted um to a certain range, which is usually set to [0, 1].

$$\mathrm{o}_{i}(t+1) = \frac{1}{1+e^{-C(t)}} \tag{2}$$

Eq. (1) describes a functional model of FCM. An FCM represents a dynamic system that evolves over time, it describes that the value of each concept is calculated by the computation of the influence of other concepts to the specific concept.

## 2.2 Inference in Fuzzy Cognitive Map

Considering it as a discrete dynamic system and calculating its final state numerically can carry out the inference process of the FCM. It is by numeric matrix operation instead of explicit IF/THEN rules[7], this is one of the main advantages of FCM.

The inference of FCM includes forward-evolved inference and backward-evolved Inference. The forward-chains of FCM to derive future states of the system it represents.. The backward-evolved inference uses the transpose of E, our FCM matrix, Et. Backward-evolved inference yields a specific concept node value that should be accompanied with a given consequence.

The forward inference process of FCM starts with a stimulus event vector. Inputting the event into the FCM. Multiplying the stimulus vector to the FCM matrix is the first in a series of such multiplications that eventually yields one of the following:

A fixed point: if the FCM equilibrium state of a dynamical system is a unique state vector, the state vector remains unchanged for successive iterations, then it is called the fixed point.

A limit cycle: if the FCM settles down with a state vector repeating in the form
$A_1 \rightarrow A_2 \rightarrow \ldots \rightarrow A_i \ldots \rightarrow A_1$
then this equilibrium is called a limit cycle.

A chaotic attractor: the FCM state vector keeps changing with each iteration repeating states are never found.

We Infer with FCM we pass state vectors X repeatedly through the FCM connection matrix W, thresholding or non-linearly transforming the result after each pass. We illustrate this by the following example: The first concept node vector be:**X**1 = (1 0 0 0 0), the connection matrix W:

$$X(t+1)=\begin{bmatrix} x_1(t+1) \\ x_2(t+1) \\ \vdots \\ x_n(t+1) \end{bmatrix}=f(WX^T(t))=f(\begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ w_{n1} & w_{n2} & \cdots & w_{nn} \end{bmatrix}\begin{bmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_n(t) \end{bmatrix}) \qquad W=\begin{bmatrix} 0 & 0 & -1 & 0 & 1 \\ 0 & 0 & 0 & -1 & 0 \\ 0 & -1 & 0 & 0 & -1 \\ -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

$X_1W$=[0,0,-1,0,1] $X_2$=f($X_1$W) (1,0,0,0,1) =$X_2$ $X_2$W=[0,0,-1,1,1] $X_3$=f($X_3$W) (1,0,0,1,1) =$X_3$ $X_3$W=[-1,1,-1,1,1] $X_4$=f($X_3$W) (1,1,0,1,1) =$X_4$ $X_4$W=[-1,1,-1,0,1] $X_5$=f($X_4$W) (1,1,0,0,1) =$X_5$ $X_5$W=[0,0,-1,0,1] $X_6$=f(X5W) (1,0,0,0,1) =$X_6$=$X_2$

So $X_2$ is a fixed point of the FCM dynamic system. This example illustrates that we can apply this kind of FCM-based forward-evolved inference approach to decision-making problems.

We can also compute backward-evolved inference by using the transpose of W, our FCM matrix, $W_t$. Backward-evolved inference yields a specific concept node value that should be accompanied with a given consequence, it does the opposite with forward-evolved inference.

## 3 Immune Algorithms

### 3.1 Natural Immune System

The natural immune system is a complex adaptive pattern-recognition system that defends the body from foreign pathogens. The main purpose of the immune system is to recognize all cells within the body and categorize those cells as self or non-self. It has dramatic and complex mechanisms that recombine the gene to cope with the invading antigens, produce the antibodies and exclude the antigens. A two-tier line of defense is in the system including the innate immune system and adaptive immune system, its basic components are lymphocytes and antibodies [8]. The lymphocyte is the main type of immune cell participating in the immune response that possesses the attributes of specificity, diversity, memory, and adaptability, there are two subclasses of the lymphocyte; T and B. each of these has its own function. The B-lymphocytes are the cells produced by the bone marrow, where each exhibits a distinct chemical structure. A B-lymphocyte can be programmed to make only one antibody that is placed on the outer surface of the lymphocyte to act as a receptor. The antigens will only bind to these receptors with which it ma In contrast, the T-lymphocytes are the cells produced by the thymus. kes a good fit [9]. By use of these T-lymphocytes, they help regulate (suppression or promotion) the production of antibodies. These receptor molecules are able to recognize disease-causing pathogens. When antigens and receptor molecules have complementary shapes, they can bind together. Once the binding ensures the recognition of the antigen, the immune response proceeds. After an antigen is recognized by immune cell receptors, the antigen stimulates the B-cell to proliferate and mature into terminal (non-dividing) antibody secreting cells (plasma cells) [10].

Based on above facts, for solving the optimization problems, the antibody and antigen can be looked as the solution and objection function, respectively.

### 3.2 Immune Algorithm

The immune algorithm (IA) is a heuristic search and optimization technique inspired by simulating the principles of natural immune system to guide their trek through a search space.

The immune algorithm is a 8-tuple (C, E, P, M, S, R, U, T), where C-coding; E-fitness function; P-initialize antibody population; M-population size; S-immune selection; R-crossover; U-mutation; T-stopping condition.

The main steps of this algorithm are as follows:

1. An initial individuals (antibodies) population are randomly formed;
2. Calculate the objective function and the affinity between antigen and antibodies and normalize the vector of the objective function;
3. The set of the cloned will suffer the crossover and mutation operation process;
4. Select the best n individuals (antibodies) with highest fitness values;
5. Clone the best n individuals (antibodies) ;
6. The fitness values of these new individuals (antibodies) are calculated.
7. Check the stopping criterion. If a termination condition is satisfied, stop the algorithm. Otherwise, go to step 2.

## 4 Goal-Oriented Analyses

Many decision support systems ask the user to identify a goal and then proceed directly to the process of finding recommendations for achieving the selected goal. The goal –oriented decision analysis starts with an expectation goal that can cause an FCM to converge to a given fixed or limit cycle attractor. The objective of finding the appropriate initial state from among a large number of possible states that an FCM can represent is essentially a search problem, which can be optimized. This turns out to be an NP hard problem in FCM's.

### 4.1 The Proposed Approach

The forward-evolved inference uses a series of vector matrix multiplication to find attractors for a given stimulus vector, while the backward-evolved inference aims to find what initial state, it does the opposite with forward-evolved inference.

The presented methodology for backward-evolved inference proposed by following four steps:

1) Define the goal state vector F;
2) Define a diagonal matrix D for calculate error. The i$^{th}$ diagonal value, $d_i \in [0,1]$,of matrix D;
3) Use IA to Optimize the objective function in Eq(3)

$$E(a,g)=(a-g)D(a-g)^T \tag{3}$$

Where E(a,g) defines the error between the target state G and the attractor A. a denotes the attractor the FCM converges to, g denotes the target state.

The essential elements of algorithm are explained as follows.

## 4.2 Explanation of the Algorithm

### 4.2.1 Antibody Encoding

In the proposed algorithm, individuals (antibodies) population is the binary coded. Each antibody is defined as a vector and consists of n variables.

Definition: $\boldsymbol{B} = [x_1 x_2 \ldots x_i \ldots x_n]$

Each antibodies can be decoded back into a candidate initial state.

### 4.2.2 Fitness Function

The search initial state from among a large number of possible states requires the definition and calculation of an objective examining function (usually an error function) when the objective function reaches a minimum that corresponds to a set of weights. When the objective function is very small, a steady state is reached.

We define the following objective function, where, *A* denotes the candidate initial state vector, is G denotes the target state.D is a n×n diagonal matrix.

$$\mathbf{E}(\boldsymbol{A},\boldsymbol{G})=\min(\mathbf{E})\ e(a,g)=(a-g)D(a-g)^T \tag{4}$$

The objective function can be used as the core of fitness function

***F (x)=I (E (x)),*** where I is an auxiliary function.

The following function ***g*** is used, the fitness function is normalized to (0, 1):

$$I(x) = \frac{1}{x+1} \tag{5}$$

### 4.2.3 Antibody Selection

In order to guarantee diversity of antibody, we use consistency-adjusting factor based on fitness selection.

Define p is selection probability of antibody, $p_f$ is selection probability based on fitness, $p_d$ is selection probability based on consistency antibody.

$$p_i = \alpha p_{fi} + (1-\alpha) p_{di} = \alpha \frac{f(i)}{\sum_{i=1}^{n} f(j)} + (1-\alpha)\frac{1}{n} e^{-\mu C_i} \tag{6}$$

Where $\alpha, \mu$ are adjusting constant, n is antibody number, $C_i$ is the consistency of antibody, its is calculated as given below.

The antibody pool is seen composed of N antibodies having M genes. For those cells marked with S={k$_1$, k$_2$, …, k$_s$}, they are alleles that come from the jth gene. From the information theory, the entropy H$_j$ (N) of the jth gene in the immune system can be computed as given below:

$$H_j(N) = \sum_{i=1}^{S} - p_{ij} \log p_{ij} \tag{7}$$

Where P$_{ij}$ is the probability that the ith allele comes out of the jth gene. By this entropy calculation, it assigns a measure of uncertainty to the occurrence or non-occurrence not of a single allele of genes, but of the whole set of alleles of genes. Note that if all alleles at jth genes are the same, then the entropy of that gene becomes zero.

The similar degreed between antibody u and antibody v:

$$ac_{uv} = \frac{1}{1+H(2)} \tag{8}$$

the consistency of antibody v:

$$C_v = \frac{1}{n}\sum_{i=1}^{n} c_{vj} \quad c_{ij} = \begin{cases} 1 & ac_{ij} > T_{ac} \\ 0 & other \end{cases} \tag{9}$$

Where: ac$_{ij}$ is the similar degreed between antibody i and antibody j, T$_{ac}$ is threshold.

### 4.2.4 Genetic Operators

The implementation of genetic operations is same as in genetic algorithms. It including the crossover operator and mutation operator requires the selection of the crossover point(s0 and mutation point(s) for each antibody under a predetermined crossover probability and mutation probability. The crossover operator provides search of the sample space to produce good solutions. The mutation operator performs random perturbations to selected solutions to avoid the local optimum.

There are many different crossover operators. In our experiments, we consider uniform crossover. Since strong relativities between weights of FCM, there is a small effect to evolution of FCM if we only change one of them. Uniform crossover generalizes this scheme to make every locus a potential crossover point. A crossover mask, the same length as the individual structures is created at random and the parity

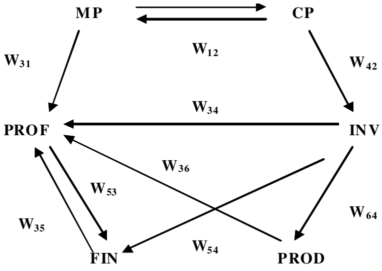


**Fig. 1.** FCM modeling

of the bits in the mask indicates which parent will supply the offspring with which bits. In natural evolution, mutation is a random process which one allele of a gene is replaced by another to produce a new individual structure. In the algorithm, Also, random mutation and roulette wheel selection are applied.

## 5 Application

To demonstrate the feasibility of the proposed method, we applied the method to Goal-Oriented Analysis. We adopted following example. The system structure is shown in Fig.1 [11]. Where MP-market position; CP-competitive position; PROF-profitability; FIN-Financing position; PROD-Productivity position; INV-Investments.

The connection matrix W of decision system and define diagonal matrix D are follows, we set the goal state vector F= [0.6, 0.55, 0.7, 0.5, 0.5, 0.6]:

$$W=\begin{bmatrix} 0 & 0.65 & 0 & 0 & 0 & 0 \\ 0.46 & 0 & 0 & 0 & 0 & 0 \\ 0.54 & 0 & 0 & 0.33 & 0.14 & -0.05 \\ 0 & 0.23 & 0 & 0 & 0 & 0 \\ 0 & 0 & -0.18 & 0.31 & 0 & 0 \\ 0 & 0 & 0 & 0.27 & 0 & 0 \end{bmatrix} \quad D = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Using Immune algorithm optimize the search for the initial state from among a large number of possible states. We set the population (antibodies) size at 50, and probability of crossover: 0.7, probability of mutation: 0.015, the maximum number of generations: 500.

The search result is follows:[0.5514,0.6103,0.7012,0.6237,0.5403,0.5223]. The average fitness among individuals in offspring with generations is shown in Fig.2. The test results show that the method is capable of Goal-Oriented Analysis.

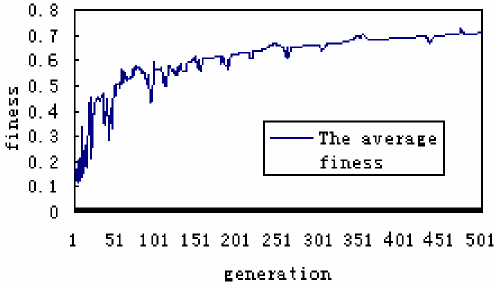


**Fig. 2.** The average fitness

## 6 Conclusion

We have developed a method for goal-oriented analysis and have discussed how immune optimization finds the appropriate initial state from among a large number of possible states. The feasibility and effectiveness of the method were illustrated. Consummating the proposed method and exploring the applying area are the direction of our future work.

## References

1. Kosko, B.: International Journal of Man-Machine Studies 24, 65–75 (1986)
2. Stylios, C.D., Groumpos, P.P.: Fuzzy Cognitive Maps: A Soft Computing Technique for Intelligent Control. In: Proc. 2000 International Symposium on Intelligent Control. Patas, pp. 97–102 (2000)
3. Liu, Z.Q., Satur, R.: Contextual fuzzy cognitive map for decision support in geographic information systems. IEEE Trans. Fuzzy Syst. 7, 495–507 (1999)
4. Satur, R., Liu, Z.Q.: A contextual fuzzy cognitive map framefwork for geographic information systems. IEEE Trans. Fuzzy Syst. 7, 481–494 (1999)
5. Pelaez, C.E., Bowles, J.B.: Applying fuzzy cognitive maps knowledge- representation to failure modes effects analysis. In: Proc. IEEE Annu. Reliability Maintainability Symp., New York, NY, January 1995, pp. 450–456 (1995)
6. Perusich, K.: Fuzzy cognitive maps for policy analysis. In: Proc. Int. Symp. Technol. Soc. Tech. Expertise Public Decisions, New York, pp. 369–373 (1996)
7. Tsadiras, A.K., Margaritis, K.: An experimental study of the dynamics of the certainty neuron fuzzy cognitive maps. Neurocomputing 24, 95–116 (1999)
8. Farmer, J.D., Packard, N.H., Perelson, A.S.: The immune system, adaptation, and, machine learning. Physica D 22, 187–204 (1986)
9. Huang, S.J.: Enhancement of thermal unit commitment using immune algorithms based optimization approaches. Electrical Power and Energy Systems 21, 245–252 (1999)
10. De Castro, L.N., Von Zuben, F.J.: The clonal selection algorithm with engineering applications. In: Proceedings of the Workshop on GECCO 2000, Las Vegas, July 8–12, pp. 36–37 (2000)
11. Carlsson, C., Fuller, R.: Adaptive Fuzzy Cognitive Maps for Hyper-knowledge Representation in Strategy Formation Process. In: Proceedings of the International Panel Conference on Soft and Intelligent Computing, Technical Univ. of Budapest, pp. 43–50 (1996), http://www.abo.fi/~rfuller/asic96.pdf

# Combination Study of Multi Fuzzy Cognitive Map

Chun-Mei Lin

Department of Computer, Shaoxing College of Arts and Sciences
Shaoxing Zhejiang 312000, China

**Abstract.** Multi-expert constructing Fuzzy cognitive map is a typical multi-expert knowledge combination problem. In this paper, we investigate the use of Dempster-Shafer evidence theory as a tool for multi-expert knowledge combination. In proposed method, we use each expert opinion as a evidence, the possible value of weight as frame of discernment, the expert's evaluation to a weight on frame of discernment as basic probability assignment, and Dempster-Shafe rule as combined basis of basic probability assignment m. Finally, the weight is given according to combined basic probability assignment. The strategy can gradually reduce the hypothesis sets and approach the truth with the accumulation of evidences, which make the result of decision more all –around and more scientific. The experimental result is shown that the method can keep exactitude information, reduce conflict factor and improve knowledge quality.

## 1 Introduction

Multi-expert constructing fuzzy cognitive map (FCM) is a typical multi-expert knowledge combination problem. Generally, the constructing FCM process is that each expert builds individual FCM, and then combines them by weight average. However, the method cannot effectively keep exactitude information, reduce conflict factor and improve knowledge quality. There is an urgent need to develop methods for multi-expert knowledge combination. Dempster-Shafer evidence theory provides solving method for the problem. In this paper, we investigate the use of Dempster-Shafer evidence theory as a tool for multi-expert knowledge combination

The paper is organized as follows. Section 2 presents the formalization representation of FCM. Section 3 presents the basic concepts of evidence theory. Section 4 presents how to use evidence theory for Multi-expert opinions combination. Section 5 applies the proposed methodology to multi-expert opinions combination. Section 6 is the conclusion and suggestions for future works.

## 2 Fuzzy Cognitive Map

FCM[4][5] is an approach to knowledge representation and inference that are essential to any intelligent system. It emphasizes the connections as basic units for storing knowledge and the structure represents the significance of system. FCM can be easily built and represent knowledge directly, And form mapped relations with the knowledge

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 332–339, 2008.

structures in the brains of the experts of this area, FCM have been used for representing knowledge and artificial inference and have found many applications, for instance, geographic information systems [6], [7], fault detection [8],, policy analysis [9], etc.

A FCM consists of nodes-concepts, each node-concept represents one of the key-factors of the system, and it is characterized by a value C$\in$ (0,1), and a causal relationship between two concepts is represented as an edge $w_{ij}$. $w_{ij}$ indicates whether the relation between the two concepts is direct or inverse.

## 3 Dempster-Shafer Evidence Theory

In this section we briefly review basic notions of Dempster-Shafer theory of evidence.

Dempster-Shafer evidence theory provides a powerfully intelligent tool for multi-expert Opinions combination. It is introduced by Dempster[1] and extended later by Shafer[2]. Dempster-Shafer theory is concerned with the question of belief in a proposition and systems of propositions. Evidence can be considered in a similar way when forming propositions, and it is concerned with evidence, weights of evidence and belief in evidence. The theory does not make any assumption concerning the way human imagination works. Simply, it describes decision-makers receiving information from different sources and evaluating to what extent the evidence that they provide is compatible or contradictory.

### 3.1 Frame of Discernment

In Dempster-Shafer theory, a problem domain is represented by a finite setΩof elements; An element can be a mutually hypothesis, an object or our case a fault. we calledΩas the frame of discernment. In the standard probability framework, all elements inΩare assigned a probability. And when the degree of support for an event is known, the remainder of the support is automatically assigned to the negation of the event.

### 3.2 Mass Functions, Focal Elements and Kernel Elements

When the frame of discernment is determined, the mass function m is defined as a mapping of the power set m: $2^{\Omega} \rightarrow [0,1]$

$$\mathrm{m}(\phi) = 0 \tag{1}$$

$$\sum_{A \subset \Omega} m(A) = 1 \tag{2}$$

The mass function m is also called a basic probability assignment function. m (A) expresses the proportion of all relevant and available evidence that supports the claim that a particular element of H belongs to the set A but to no particular subset of A. In engine diagnostics, m(A) can be considered as a degree of belief held by an observer regarding a certain fault; different evidence can produce different degrees of belief with respect to a given fault. Any subset A of Ω such that m(A) > 0 is called a focal element; the union of all focal element C $=\cup_{m(A)\neq 0}A$ is called a kernel element of mass function m in the frame of discernment.

### 3.3 Belief and Plausibility Functions

The belief function Bel is defined as:

Bel : $2^{\Omega}\rightarrow$ [0,1] $\forall A\subset\Omega$

$$\mathrm{PI(A)}=1-\mathrm{Bel}(\overline{\mathrm{A}})=\sum_{B\subseteq\Omega}m(B)-\sum_{B\subset\overline{A}}m(B)=\sum_{B\cap A\neq\Phi}m(B) \tag{3}$$

The belief function Bel(A) measures the total amount of probability that must be distributed among the elements of A; it reflects inevitability and signifies the total degree of belief of A and constitutes a lower limit function on the probability of A.

The plausibility function Pls and double function Dou are defined as:

Pl: $2^{\Omega}\rightarrow$ [0,1]

$$\mathrm{PI(A)}=1-\mathrm{Bel}(\overline{\mathrm{A}}),\mathrm{Dou(A)}=\mathrm{Bel}(\overline{\mathrm{A}}) \tag{4}$$

The plausibility function Pl(A) measures the maximal amount of probability that can be distributed among the elements in A; it describes the total belief degree related to A and constitutes an upper limit function on the probability of A. it describes the total belief degree related to A and constitutes an upper limit function on the probability of A.

### 3.4 Evidence Combination

Let $\mathrm{Bel}_1$ and $\mathrm{Bel}_2$ be two belief functions in the same frame of discernment, then the corresponding basic belief assignment are $m_1$ and $m_2$ based on information obtained from two different information sources in the same frame of discernment Ω, focus elements are X1, X2, …, Xk and Y1, Y2, …, Yk, if Xi∩Yj=A, $X\subset\Omega$, then $m_1(X_i)m_2(Y_j)$ is the probability assignment to A, The total belief of A is:

$$\sum_{X_i\cap Y_j=A} m_1(X_i)m_2(Y_j),\ A\neq\phi$$

when $A=\phi$, $\sum_{X_i\cap Y_j=\phi} m_1(X_i)m_2(Y_j)$ is on the belief of void setφ. We have the rule of evidence combination[3].

$$\mathrm{m(A)}=m_1\oplus \mathrm{m}_2=\begin{cases}0 & A=\Phi\\ \frac{1}{1-\mathrm{k}}\sum_{X\cap Y=A} m_1(X)m_2(Y) & A\neq\Phi\end{cases} \tag{5}$$

Where $\mathrm{k}=\sum_{X\cap Y=\phi} m_1(X)m_2(Y)$, K represents a basic probability mass associated with conflicts among the sources of evidence. It is determined by summing the products of mass functions of all sets where the intersection is null. K is often interpreted as a measure of conflict between the sources. The larger the value of K is, the more conflicting are the sources, and the less informative is their combination.

The produced function m =m1 $\oplus$ m2 is also a mass function in the same frame of discernment $\Omega$, it represents the combination of m1 and m2 and carries the joint information from the two sources.

In the case of n mass functions $m_1, m_2, \ldots, m_n$ in $\Omega$, according to rule of evidence combination:

$$\mathrm{m(A)} = m_1 \oplus \mathrm{m}_2 \oplus \ldots \oplus m_n = \begin{cases} 0 & A=\Phi \\ \dfrac{\sum\limits_{A_1 \cap A_2 \cdots \cap A_n = A} \prod\limits_{i=1}^{n} m_i(A_i)}{1\text{-k}} & A \neq \Phi \end{cases} \tag{6}$$

Where k= $\sum\limits_{A_1 \cap A_2 \cdots \cap A_n = \Phi} \prod\limits_{i=1}^{n} m_i(A_i)$

## 4 Multi-expert Opinions Combination

According to the formulized definition of FCM, experts' opinions are reflected on the estimate of the degree of the cause that is between nodes in the referred concept set, namely weight estimate. In the construction of FCM, multi-experts' opinions combination is represented as the combination of the corresponding elements in the connection matrix provided by experts. Then each expert's estimate of some cause relation can be regarded as evidence. The possible values of the affection degree of the cause relation between concepts form a frame of discernment. The combined probability assignment function is regarded as the evidence of last weight integration.

A FCM equals the code of experts' knowledge; In general, because of experts' different preferences and knowledge structures, the understandings about the problem may be different. Such as, different experts differ in how they assign causal strengths to edges and in which concepts they deem causally relevant. There is a requirement to build a selection rule of concept set and to enact a standard of cause effect degree before FCM combination.

Definition 1: Connection Matrix Standardization

Suppose there are n experts, the FCM of each expert's is established according to their own experiences and knowledge. The connection matrices of n experts' are F1, F2, …, Fn. The union (m) of all experts' concepts is regarded as a set of concept. The connection matrices of experts' are expanded to m×m, and we fill the row or column absent of concept nodes with 0. The process is called the standardization of connection matrix.

The general process of combining multi-experts' FCM with evidence theory is as follows:

1. A frame of discernment is firstly defined, it translates the research of proposition into the research of a set.
2. Basic probability assignments are established according to evidence.

3. Basic probability assignment functions are combined according to the combination rule of evidence theory, then the target type is determined by the rule of belief evaluation.
4. Applying weighted average on all elements of the frame according to the integrative basic probability assignment function

### 4.1 Building of Discernment Frame

The selection of frame of discernment depends upon our knowledge, cognition and what we know and want. In application of FCM, expert estimates the weight using linguistic weight. Their values are usually nothing, very weak, weak, medium, strong, very strong.

Example 1:
(none, very weak, weak, strong, very strong, extremely strong) →{0, 0.2, 0.4, 0.6, 0.8, 1}

Example 2:
(none, weak, strong, extremely strong) →{0, 0.4, 0.6, 1}

The possible values of weight form a frame of discernment, which is defined by the demand of accuracy.

We can define a frame of discernment according to the example above.

$\Omega=\{0,\ 0.2,\ 0.4,\ 0.6,\ 0.8,\ 1\}$

Or: $\Omega=\{0,\ 0.4,\ 0.6,\ 1\}$

### 4.2 Building of Mass Function

According to the experience and knowledge, each expert makes a basic probability assignment function m(also called the mass function m ) for every element of the connection matrix in a frame of discernment. Suppose there are n experts, we can gain n basic probability assignment functions: $m_1, m_2, \ldots, m_n$ .

N experts' evaluating a weight in the frame of discernment can capture a matrix form: The matrix M is as follows:

$$M = \begin{bmatrix} m_{11} & m_{12} & \cdots & m_{1k} \\ m_{21} & m_{22} & \cdots & m_{2k} \\ & & \vdots & \\ m_{n1} & m_{n2} & \cdots & m_{nk} \end{bmatrix}$$

Where each row in matrix M represents the evaluation of $i^{th}$ expert; Each column of matrix M represents the evaluation that n experts get after evaluating the $j^{th}$ element of the frame of dscernment. $m_{ij}$ denotes the $i^{th}$ expert's probability assignment of the $j^{th}$ element of the frame of discernment Ω.

The result that the $i^{th}$ expert estimates a weight in the frame of discernment is a fuzzy value. A basic probability assignment function $m_i$ is produced by solving membership of the fuzzy value.

For example, solveing m, when a fuzzy value is 0.48, the membership function is defined as follows:

Fig.1. the six membership functions corresponding to each one of the six linguistic variables the fuzzy number

$$\mu_0(x) = \begin{cases} (0.2 - x)/0.2 & 0 < x \le 0.2 \\ 0 & other \end{cases}$$

$$\mu_{0.2}(x) = \begin{cases} (0.2 - x)/0.2 & 0 < x \le 0.2 \\ (0.4 - x)/0.2 & 0.2 < x \le 0.4 \\ 0 & other \end{cases}$$

$$\mu_{0.4}(x) = \begin{cases} (x - 0.2)/0.2 & 0.2 < x \le 0.4 \\ (0.6 - x)/0.2 & 0.4 < x \le 0.6 \\ 0 & other \end{cases}$$

$$\mu_{0.6}(x) = \begin{cases} (x - 0.4)/0.2 & 0.4 < x \le 0.6 \\ (0.8 - x)/0.2 & 0.6 < x \le 0.8 \\ 0 & other \end{cases}$$

$$\mu_{0.8}(x) = \begin{cases} (x - 0.6)/0.2 & 0.6 < x \le 0.8 \\ (1.0 - x)/0.2 & 0.8 < x \le 1.0 \\ 0 & other \end{cases}$$

$$\mu_{1.0}(x) = \begin{cases} (1.0 - x)/0.2 & 0.8 < x \le 1.0 \\ 0 & other \end{cases}$$

According to formula above:$\mu_{0.4}(0.48)=0.6$, $\mu_{0.6}(0.48)=0.4$

We can obtain the base probability assignment m=[0,0,0.6,0.4,0,0]

### 4.3 Evidence Combination

Firstly, the conflict between the expert's opinions is calculated with $k = \sum_{X \cap Y = \phi} m_1(X) m_2(Y)$.

If combination condition is satisfied, then the combinative base probability assignment is calculated according to formula bellow:

$$\text{m(A)} = m_1 \oplus \text{m}_2 = \begin{cases} 0 & A = \Phi \\ \dfrac{1}{1 - \text{k}} \sum_{X \cap Y = A} m_1(X) m_2(Y) & A \ne \Phi \end{cases}$$

### 4.4 Calculate Integrated Weight

The integrated weight *w* is defined as:

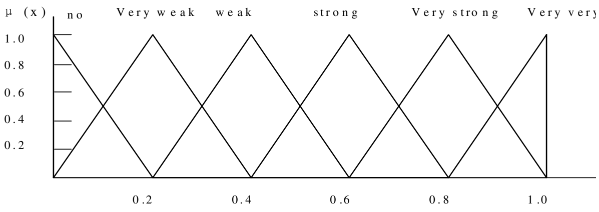
$$w = \sum_{j=1}^{m} \frac{a_j}{\sum_{j=1}^{m} a_j} \theta_j \tag{7}$$

Where $a_j$ is the base probability assignment of the *j*th state, $\theta_j$ is the *j*th state value of the frame of discernment.

## 5 Application

To demonstrate the feasibility of the proposed method, we applied the proposed method to the combination of three experts' opinions.

We define a frame of discernment:



Ω={0, 0.2, 0.4, 0.6, 0.8, 1}

three experts give judgment to the cause affection degree of Ci and Cj in concept set{C1, C2, …Cn}, see table 1.

**Table 1.** Expert knowledge

| State expert | 0 | 0.2 | 0.4 | 0.6 | 0.8 | 1 |
|---|---|---|---|---|---|---|
| Expert 1 | 0 | 0 | 0.7 | 0.3 | 0 | 0 |
| Expert 2 | 0 | 0.1 | 0.8 | 0. | 0 | 0 |
| Expert 3 | 0 | 0 | 0.9 | 0.1 | 0 | 0 |

According to formula $k = \sum_{X \cap Y = \phi} m_1(X) m_2(Y)$, we get k=0.41.

The combinative result of Expert 1 and expert 2 according to Eq (7) is shown in table 2.

**Table 2.** Experts knowledge combination (1)

| State expert | 0 | 0.2 | 0.4 | 0.6 | 0.8 | 1 |
|---|---|---|---|---|---|---|
| Expert 1 | 0 | 0 | 0.7 | 0.3 | 0 | 0 |
| Expert 2 | 0 | 0.1 | 0.8 | 0.1 | 0 | 0 |
| Combination result 1 | 0 | 0 | 0.95 | 0.05 | 0 | 0 |

Again, according to $k = \sum_{X \cap Y = \phi} m_1(X) m_2(Y)$, we get k=0.14.

The combinative result of expert 1, expert 2 and expert 3 according to Eq (7) is shown in table 3.

**Table 3.** Experts knowledge combination (2)

| State Expert | 0 | 0.2 | 0.4 | 0.6 | 0.8 | 1 |
|---|---|---|---|---|---|---|
| Result 1 | 0 | 0 | 0.95 | 0.05 | 0 | 0 |
| Expert 3 | 0 | 0 | 0.9 | 0.1 | 0 | 0 |
| Combination result | 0 | 0 | 0.994 | 0.0058 | 0 | 0 |

The result of three experts' conbination can be seen from table 3. The base probability assignment is 0.994 when state value is 0.4 and m(0.6) is 0.0058.

Using Eq (9) to solve the integrated weight according to the combined base probability assignment function m.

Based on the example above, we get $w_{ij}$ :$w_{ij}$=0.4*0.994+0.6*0.0058=0.40108.

## 6 Conclusion

We have developed a method for Multi-Expert Opinions Combination Based on Evidence Theory. . In the method, we use multi-expert knowledge as evidence, the possible value of weight as frame of discernment, expert's evaluation to a weight on frame of discernment as basic probability assignment, and Dempster-Shafe rule as combined basis of basic probability assignment m. Finally, the weight is given according to combined basic probability assignment. The strategy can gradually reduce the hypothesis sets and approach the truth with the accumulation of evidences, which make the result of decision more all–around and more scientific. Consummating the proposed method and exploring the applying area are the direction of our future work.

## References

1. Dempster, A.P.: Upper and lower probabilities induced by a multi-valued mapping [J]. Ann. mathematical Statistics, 325–329 (1967)
2. Shager, G.: Constrative Probability. Synthese 48, 1–60 (1981)
3. Shafer, G.: A Mathematical Theory of Evidence. Princeton University Press, New Jersey (1976)
4. Kosko, B.: International Journal of Man-Machine Studies 24, 65–75 (1986)
5. Liu, Z.Q., Satur, R.: Contextual fuzzy cognitive map for decision support in geographic information systems. IEEE Trans. Fuzzy Syst. 7, 495–507 (1999)
6. Liu, Z.Q., Satur, R.: Contextual fuzzy cognitive map for decision support in geographic information systems. IEEE Trans. Fuzzy Syst. 7, 495–507 (1999)
7. Satur, R., Liu, Z.Q.: A contextual fuzzy cognitive map framefwork for geographic information systems. IEEE Trans. Fuzzy Syst. 7, 481–494 (1999)
8. Pelaez, C.E., Bowles, J.B.: Applying fuzzy cognitive maps knowledge- representation to failure modes effects analysis. In: Proc. IEEE Annu. Reliability Maintainability Symp., New York, NY, January 1995, pp. 450–456 (1995)
9. Perusich, K.: Fuzzy cognitive maps for policy analysis. In: Proc. Int. Symp. Technol. Soc. Tech. Expertise Public Decisions, New York, pp. 369–373 (1996)

# A Hermeneutic Approach to the Notion of Information in IS

Su-Fen Wang

Glorious Sun School of Business & Management, Donghua University
Shanghai 200051, China

**Abstract.** In the field of information systems (IS), 'information' is probably the most important and fundamental notion. And yet, existing studies of it seem inadequate and lacking of sufficient depth, and further work is therefore desperately in need. We adopt Hermeneutics to approach the essence of information. We describe how Hermeneutics might enable us to look at the mechanism whereby information is created and information flow takes place, and explain implications this approach might have to requirement identification in IS.

## 1 Introduction

Arguably 'information' is probably the most fundamental and the most important concept for information systems and information science [1]. Boland maintains that since the very beginning of the emergent discipline of information systems, research on information systems has suffered from the elusive nature of information and the lack of methods and techniques for handling the essence of information, and points out that how well this problem is addressed has profound impact to all aspects of information systems – the research, development and use of information systems, both in theory and in practice [1]. This is one hand. On the other hand, there are various forms about such key concepts as *information, data and meaning* ([2],[3],[4],[5],[6]). These concepts and terminology are often isolated, disjoint, and form an often contradictory amalgam of knowledge and cause confusion in research. Therefore, 'How to understand and to deal with the essence of information' would seem to have become an important and tough problem for the discipline of IS.

We aim to tackle this problem. We put forward a new approach to this problem by using results of 'information philosophy' in Section 2. We suggest and explain the general process of information realization based on Hermeneutics in Section 3. In the final section, we give some concluding remarks about the work to be presented in this article and explain implications this approach might have to requirement identification.

## 2 A New Perspective on the Essence of Information

We suggest using results of research in 'information philosophy' ([7], [8]) in recent years thereby to formulate a new perspective for approaching the problem of 'information'.

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 340–345, 2008.

The new perspective is to understand and engage the essence of information from 'why information is called information in the first place' through investigating the relationship among information, data and meaning. In particular, new meaning is created through a specific type of activities and behavior, called *interpretation*. That is, we observe that in order to understand the essence of information we adopt a perspective that information in IS can only be created through practice and being engaged through interpretation. This is the pivotal point of the ideas that we develop in the paper.

We believe that information is carried by non-empty, well-formed and meaningful data [8], and an information system is a social system making use of IT. Thus the relationship among information, data and meaning can only be explored through communication and negotiations between humans. It is conducted within the never-ending cycle of '*information* is carried/embodied/projected by *data*; *meaning* is created from information through *interpretation* of data; then further information may be created due to the *intention* of a human agent, which is again carried by *data*' Through communication and negotiations between them, people obtain understanding of the world around them and of themselves. So, 'why information is called information in the first place' must be considered form the viewpoint of human's existence.

The process above is that of information (impact) realization, and the associated mechanism is that of information realization. On the one hand this process captures the relationship between information, data and meaning. On the other hand, this process is accomplished through the interaction between the three. We believe that information is independent of informees (the receivers of information), borrowing Floridi's term [8]. But we also believe that the impact of information, which is concerned with the reason why a piece of information can be seen as such, namely due to its capability of informing, can only be materialised through the interaction between the three, i.e., information, data and meaning. This entails the involvement of human agents within the process, or the interpretation/creation of meaning. Information realization is concerned with how people use information, and how information supports people who need information. Therefore the process of information realization becomes a process of meaning interpretation and realization.

The informing process through accessing information is that of interpreting the meaning of information for the informee in the sense that what the information means to him/her. It would seem that this has not been adequately addressed. Furthermore, exploring meaning would seem a basic problem for hermeneutics. In the sections that follow, we will put forward a proposal on how a mechanism for exploring meaning might look like by drawing on hermeneutics.

## 3 A Hermeneutic Approach to the Problem of Information in IS

### 3.1 The General Process of 'Information Realizing' Mechanism

Hermeneutics is the study of interpretation. Hermeneutics emerged as a concern with interpreting ancient religious texts and has evolved to address the general problem of how we give meaning to what is unfamiliar and alien([9],[10], [11]).

In the context of IS, data, information and meaning are in a state of co-existence. Information is borne by data, and meaning is created due to reception of information through looking at data or interpreting data. Thus these data are in the position of the target, i.e., 'text' in Hermeneutics.

We consider information systems as social systems that are technically and technologically implemented, so, we adopt Ricoeur's Hermeneutics as the theoretical foundation for our investigation into the mechanism that enables the realization of information and information flow within the context of information systems([10]).

Ricoeur combines ontological Hermeneutics with methodological and epistemological Hermeneutics through linking Hermeneutics with the text theory.

The general process of information and information flow realization (see Fig.1) may be seen as having three stages, namely the Semantics Layer, Reflection Layer, and Ontological Layer. Each of the layers is connected with the 'text' (i.e., data) of the information system. The transformations between the three layers embody those between objective meaning (in the sense of being independent of the receiver of information), inter-subjective meaning and subjective meaning.

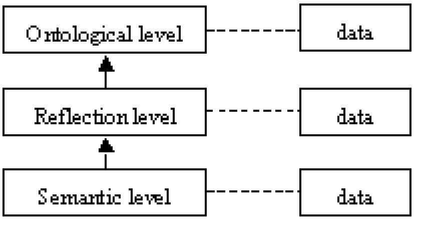


**Fig. 1.** The general process of information realization

### 3.2 An Analysis of Various Elements in the 'Information Realizing' Process

#### 3.2.1 Data Analysis

With Hermeneutics, data in information systems are read and interpreted as texts. We give data here slightly different characteristics from those that appear in more 'general' research of information systems([6], [8], [9]).

We think that data links information and meaning, which enables the communication between people. Through communication, people acquire self-understanding. Thus data should have the following characteristics:

1. Data are fixed life expressions by being written. They have multiple meanings and multiple layers of meaning. There are literal meaning, sender's meaning, hidden and latent meaning produced by various factors, such as the multiple traits of literal meaning, the knowledge background and psychological factors of the sender and so on.
2. There is a dialectic relation between the sender's meaning and the meaning that may be seen as inherent to the data([14], [15]).
3. Meaning created through information carried by data and the relevance of data is derivative from the dialectic relation between data and its receiver.
4. Data is not limited by their direct references; data enable people to enter a possible world from a given one, i.e., the data world.

Therefore, the process of information realization is a process of interpretation of multiple layers of meaning and that of realization of multiple meanings. This in turn enables data to have their complex characteristics as just discussed.

### 3.2.2 Semantic Analysis

The analysis of the information content of data, through interpreting the data, we can obtain objective information content carried by the data. The objective information content is taken as the meaning that the sender of the data wishes the data to carry. So 'objective' here means being independent of the receiver of the data. Data may have various meanings, such as the literal meaning, which may in turn refer to a particular event.

Literal meaning is the direct and basic meaning, and the others are indirect, second or metaphoric meaning. These indirect meanings are nested within the direct meaning. This is similar in a way to information nesting [12].

We begin to interpret data that have multi-stipulations. But every kind of interpretation is based on its own frame of reference in order to seek agreement with the rich and multi-vocal meanings of data. The interpretation process of data is illustrated in Fig. 2.

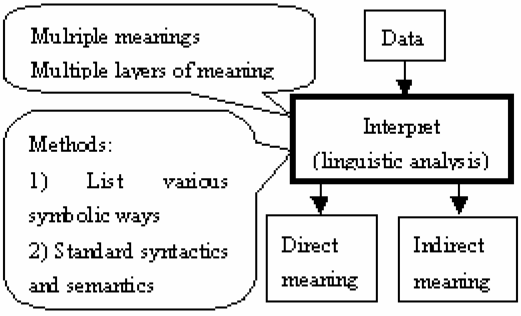

Mulriple meanings
Multiple layers of meaning
Data
Interpret
(linguistic analysis)
Methods:
1) List various symbolic ways
2) Standard syntactics and semantics
Direct meaning
Indirect meaning

**Fig. 2.** The process of linguistic analysis

### 3.2.3 Reflection Layer

The information forming process embodies the communication between people by means of the inter-relationship of data, information and meaning. Its goal is for people to achieve understanding of themselves by communicating with one another.

Thus our interpretation of data is not just the understanding of the information content that is carried by the data, but also the meaning of the sender of the data. The purpose of this is, through understanding the sender's meaning, to ascertain what world we ourselves are in, and make sure of what 'I' am, and what I should do. This is self-understanding, to achieve which there has to be a process of reflection.

Reflection is of course self-reflection, and not a concrete reflection on a particular event. Reflection is a process of transforming the 'otherness' of the data into an 'utterance event' for me. The receiver's 'utterance event' is a new event, that is, it is not the repetition of the ''utterance event' that created the data in the first place, but is new creation according to the requirements of 'speaking'. This way, the interpretation of reflection is completed. Thus, self-understanding is realized through reflection.

Reading links two incidents of speaking: data as utterances, and reading as new utterances. Ricoeur makes use of Gardmer's 'fusion of horizons' to refer to the widening of the understanding of the subject after she/he has entered the world of data.

We place data at the position of a production medium. Through the interpretation cycle, more meaning is obtained; and through 'fusion of horizons', self-understanding is achieved.

Reflection process is completed through reading data and conversing with data, and reading through 'fusion of horizons' and game-playing.

### 3.2.4 Ontological Layer

After reflection, self-understanding comes into being according to the form in which it can exist, and it creates new data. This is not an end, but the beginning of a new cycle. This process of information realization and information flow constitute a basis of exchange between human being.

## 4 Concluding Remarks - Implications to Requirements Identification for Information Systems

Through semantic interpretation of the semantic layer, the receiver obtains the information content of the data sent by the sender. Much of the information content exists in the form of being implied and implicit, through obtaining which the receiver obtains her/his understanding of the sender. Through assimilation via reflection, the receiver strives to find the way to further understanding her/himself, namely to make something 'alien' to be of his/her own. Then on the ontological layer, the receiver expresses his/her own utterance with new data. Through such a never-ending cycle, human exchange is achieved, which in turn enables us to increasingly understand ourselves (see Fig. 3).

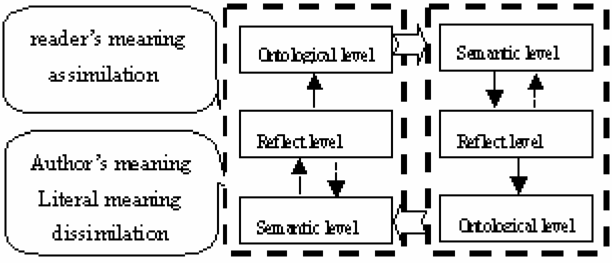


**Fig. 3.** The communication pattern between subjects

Our work along this line seems to have implications for the identification of requirements in IS. There are three problems concerning requirements of IS, namely,

- The content of IS requirements by the user
- How required information is identified, and
- How data that an IS will actually store and process are identified.

These can now be looked at within such a never-ending Hermeneutic cycle. The content of information requirement from the point of the view of the user would now be what is needed for her/him to understand her/himself in the context of using an IS to approach and complete her/his tasks and the meanings that are subsequently produced. The required information should be identified through the stages of semantic understanding, reflection and ontological realization. Finally, the data that an IS processes should be among the original set of data and the new data. These can be seen in Fig 4. To work out the details of how a mechanism for analyzing information and information flow within the context of IS would require much more work and it is therefore beyond the scope of this paper.

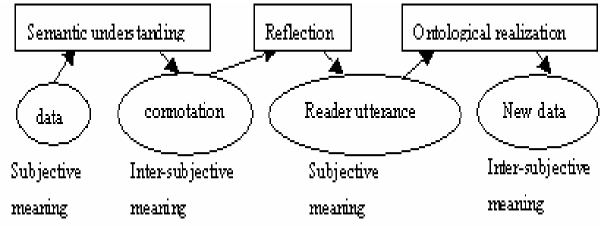


**Fig. 4.** How to identify the required information through the stages of semantic understanding, reflection and ontological realization

## References

1. Boland Jr., R.J.: The Information of Information Systems. In: Boland Jr., R.J., Hirschheim, R.A. (eds.) Critical Issues in Information Systems Research, ch.14, pp. 363–379. John Wiley & Sons LTD, Chichester (1987)
2. Checkland, P., Scholes, J.: Soft Systems Methodology in Action. Wiley, Chichester (1990)
3. Checkland, P., Holwell, S.: Information, Systems and Information Systems: making sense of the field. John Wiley & Sons Ltd, Chichester (1998)
4. Davis, G., Olsont, M.: Management Information Systems: Conceptual Foundations, Structure and Development, 2nd edn., p. 200. McGraw-Hill, New York (1985)
5. Stamper, R.: Organisation Semiotics. In: Mingers, J., Stowell, F. (eds.) Information Systems: an Emerging Discipline?, pp. 267–283. McGraw-Hill Publishing Co., New York (1997)
6. Mingers, J.: Information and meaning: foundations for an intersubjective account. Info. Systems J. 5, 285–306 (1995)
7. Floridi, L.: Philosophy of Computing and Information. Blackwell, Malden (2004)
8. Floridi, L.: Is Information Meaningful Data? Philosophy and Phenomenological Research 70(2), 351–370 (2005)
9. Hirschheim, R., Klein, H.K., Lyytinen, K.: Information systems development and data modeling. Cambridge University Press, Cambridge (1995)
10. Ru-Lun, Z.: The inquire of meaning: contemporary western hermeneutics. Liaoning Publishing (1985)
11. Gadamer, H.: Philosophical Hermeneutics. University of California Press, Berkeley (1976)
12. Dretske, Fred, I.: Knowledge and the Flow of Information. MIT Press, Cambridge (1981)

# Access Control Labeling Scheme for Efficient Secure XML Query Processing

Dong Chan An and Seog Park

Department of Computer Science & Engineering, Sogang University
C.P.O. Box 1142, Seoul Korea 100-611
{channy, spark}@sogang.ac.kr

**Abstract.** Recently XML has become an active research area. In particular, the need for an efficient secure access control method of dynamic XML data in a ubiquitous data streams environment has become very important. In this paper, we propose the access control labeling scheme for efficient secure query processing under dynamic XML data streams. The proposed labeling scheme supports the process of dynamic XML data without re-labeling existing labels and secure query processing. We have shown that our approach is efficient and secure through experiments.

**Keywords:** Labeling Scheme, Access Control, XML.

## 1 Introduction

XML has become a widely popular standard for representation and exchanging data over the Internet. Query language like XPath and XQuery are developed by W3C group for XML data. The efficient secure processing of XPath or XQuery is an important research topic. Since logical structure of XML is a tree, establishing a relationship between nodes such as parent-child relationship or ancestor-descendant relationship is essential for processing the structural part of the queries. For this purpose many proposals have been made such as structural indexing and nodes labeling. Relatively little work has been done to enforce access controls particularly for XML data in the case of query access control. Moreover, the current trend in access control within traditional environments has been a system-centric method for environments including finite, persistent and static data. However, more recently, access control policies have become increasingly needed in continuous data streams [2], and existing access control models and mechanisms cannot be adequately adopted on data streams [6].

The rest of this paper is organized as follows. Section 2 presents related work in the area of XML access control and labeling scheme. Section 3 introduces the algorithm and techniques of the proposed method. Section 4 shows the experimental results from our implementation and shows the processing efficiency of our framework. Our conclusions are contained in Section 5.

## 2 Related Work

The traditional XML access control enforcement mechanism [3], [4], [7], [8] is a view-based enforcement mechanism. The semantics of access control to a user is a

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 346–353, 2008.

particular view of the documents determined by the relevant access control policies. It provides a useful algorithm for computing the view using tree labeling. However, aside from its high cost and maintenance requirement, this algorithm is also not scalable for a large number of users.

XML data has a hierarchical structure and the required capacity might be very huge. A method that can take XML data into appropriate fragmentation so as to process it in pieces is consequently needed for the small memory of a mobile terminal to manage massive XML data [1], [7]. When XML streams data, which is generated under a sensor network, the data is structurally fragmented and transmitted and processed in XML piece streams, the efficiency of memory and the processing time of mobile terminals can be reconsidered. Moreover, when certain data is updated in an XML data streams, not the whole XML data but only the changed fragment needs to be transmitted, taking advantage of a reduced transmission cost.

The recent Hole-Filler Model [6], [4] has been proposed as a method that fragments XML data structurally. XFrag [5] and XFPro [9] proposed an XML fragmentation processing method adopting the Hole-Filler Model. Nonetheless, this method has problems of late processing time and waste of memory space due to additional information for the Hole-Filler Model.

The labeling technique of [13] has shown up as a consequence of the appearance of the dynamic XML document. This technique is typical of the prime number labeling scheme applied to information which rarely affects other labels. This technique assigns a label for each node, a prime number, to represent the ancestor-descendant relation and is designed not to affect the label of other nodes when updating the document. However, since it searches a considerable range of the XML tree again and re-records updated order information during the updating process, it presents a higher updating cost.

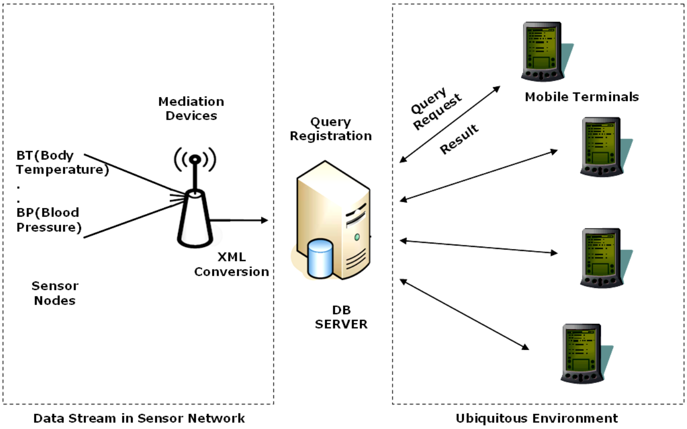
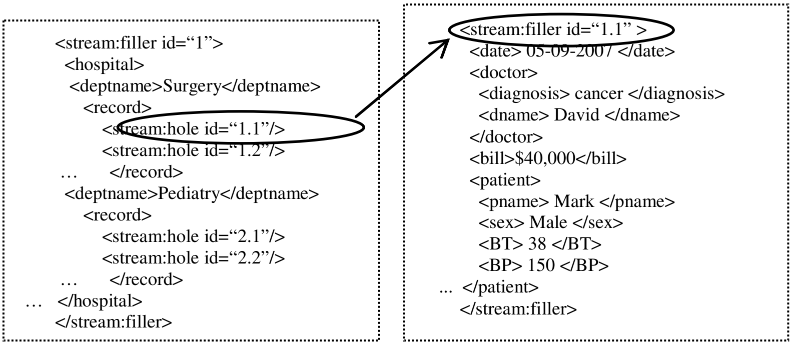


**Fig. 1.** Query Processing of Mobile Terminal under XML Data Streams Environment

## 3 Proposed Method

The proposed environment of the dynamic role-based prime number labeling scheme is shown in Fig. 1. It depicts medical records that need accurate real-time query answers by checking the authority and the role of valid users via access control policy when a query is requested. First of all, considering the characteristics of the proposed environment, the fragmentation of the XML document, problems such as low processing time and waste of memory space needed due to additional information for the Hole-Filler Model in existing XFrag [5] is minimized as shown in Fig. 2. This means that information such as tag structure is no longer needed because the order of XML documents no longer needs to be considered.

After a fragmenting of XML data streams, proper role-based prime number label for accessible roles' product are assigned to nodes of the medical records XML document as referring to Table 1. Since roles are limited in any organization, it is possible to represent roles with a prime number and a prime number is expansible.



```
<stream:filler id="1">
  <hospital>
   <deptname>Surgery</deptname>
     <record>
        <stream:hole id="1.1"/>
        <stream:hole id="1.2"/>
 ...     </record>
  <deptname>Pediatry</deptname>
     <record>
        <stream:hole id="2.1"/>
        <stream:hole id="2.2"/>
 ...     </record>
...  </hospital>
 </stream:filler>
```

```
<stream:filler id="1.1" >
 <date> 05-09-2007 </date>
 <doctor>
   <diagnosis> cancer </diagnosis>
   <dname> David </dname>
 </doctor>
 <bill>$40,000</bill>
 <patient>
   <pname> Mark </pname>
   <sex> Male </sex>
   <BT> 38 </BT>
   <BP> 150 </BP>
... </patient>
 </stream:filler>
```

**Fig. 2.** Partial Fragmentation in XML Data Streams

**Table 1.** Role-Based Prime Number Table

| Role | Role-Based Prime Number |
|---|---|
| Patient | 2 |
| Doctor | 3 |
| Researcher | 5 |
| Insurer | 7 |

### 3.1 Proposed Labeling Scheme

**Labeling Notation.** The labels of all nodes are constructed by four significant components (*l, L1, L2, and L3*), which are unique.

1. *Level component (l)* – It represents the level of node in the XML document. The level of tree from root to leaf is marked but the level of root is null.
2. *Ancestor label component (L1)* – The component that succeeds to the label of parent node, which eliminates the level component from a parent node label, is inherited.
3. *Self label component (L2)* – It represents the self label of node in the XML document.
4. *Prime number product component (L3)* – It represents the prime number product of accessible role for node.

The Fig. 3 is labeled XML tree by applying the labeling scheme.

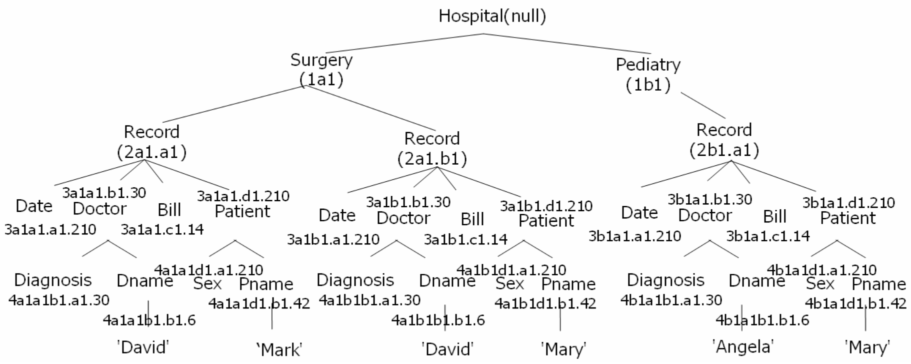


**Fig. 3.** Proposed Labeling of Medical Records XML Document

**Labeling Scheme.** The Labeling for an XML document is following.

*l L1.L2.L3*

1. The root node is the first level. Because it does not have a sibling node and parent node, root node is labeled with "null"
2. The first child of root node is labeled with label ($N_1$), "*1a1.L3*". The second child of root node is labeled with label($N_2$), "*1b1.L3*". Because parent node's label(L1) is "null", *l* and L2 is concatenated. The third component *(L3)* is optional in this level. If the third component is all roles' accessible, third component is able to omit.
3. The first child of second level $N_1$ is labeled with label ($NN_1$), "*2a1.a1.L3*". The second child of second level $N_2$ is labeled with label ($NN_2$), "*2b1.a1.L3*". Because parent node's label (*L1*) is inherited. The third component *(L3)* is labeled with prime number product for node's accessible roles.
4. The first child of third level $NN_1$ is labeled with label ($NNN_1$), "*3a1a1.a1.L3*". The second child of third level $NN_2$ is labeled with label ($NNN_2$), "*3b1a1.a1.L3*". Because parent node's label (*L1,L2*) is inherited.
5. The first child of third level $NNN_1$ is labeled with label, "*4a1a1a1.a1.L3*". The second child of third level $NNN_2$ is labeled with label, "*3b1a1.a1.L3*". Because parent node's label (*L1,L2*) is inherited.
6. The third component (*L3*)*,* prime number product of accessible role for node is generated by following.
   - Date (210) : product of (accessible role prime numbers) 2,3,5,7
   - Doctor (30) : product of 2,3,5
   - Bill (14) : product of 2,7
   - Diagnosis (30) : product of 2,3,5

### 3.2 Query Processing by 'Role-Based Prime Number'

The proposed security system's architecture is shown in Fig. 4. The query processing procedure in Fig. 4 can be considered in two steps. The role check is done in Step 1 using the ‘Role-Based Prime Number Table’ and final access authority is checked at Step 2 using the ‘Role Privacy Table’. Once a query from a mobile terminal is registered, access authority is checked at Step 1 by checking the prime number of the query terminal node. That is, access to Step 1 is permitted when the remainder of the *L3*(product of accessible role prime numbers) divided by the role of user becomes zero. Accessibility is finally checked at Step 2 referring to the ‘Role Privacy Table’ of Table 2. Moreover, query access is rejected by ‘denial-takes-precedence’ [11].

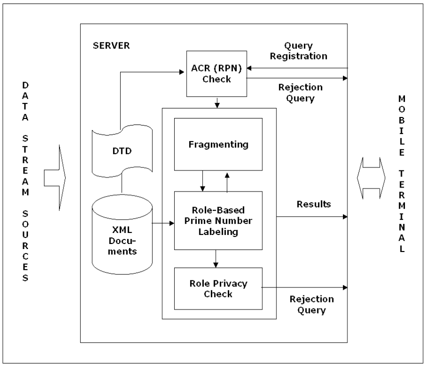


**Fig. 4.** Proposed Security System

**Table 2.** Role Privacy Table

| Department | Record | Role | | | |
|---|---|---|---|---|---|
| | | Patient | Doctor | Insurer | Researcher |
| Sugery<br>(a1) | a1.a1 | Mark | David | ING | - |
| | a1.b1 | Mary | David | AIG | - |
| | … | … | ... | … | - |
| Pediatry<br>(b1) | b1.a1 | Mary | Angela | AIG | - |
| | … | … | ... | … | - |

**Example 1.** (predicate + positive access control + negative access control)

(1) //record/patient[pname=Mark]

(2) "Angela" with role of doctor requests a query

- step1, terminal node pname=Mark’s is verified : ‘4a1a1d1.b1.42’
- Angela’s role is doctor(3), 42%3=0, access is permitted
- step2, Only prefix label ‘b1a1’ is permitted for Angela by ‘Role Privacy Table’
- [pname=Mark] is ‘4a1a1d1.b1.42’, access rejected
- access to step1 permitted, access to step2 rejected. Finally, query access rejected

**Example 2.** (negative access control)

(1) //record/patient/pname

(2) one with role of researcher requests a query

- step1, terminal node pname’s label is verified : *.*.42
- researcher’s role is 5, 42%5≠0, access rejected. Finally, query rejected

As shown in Example 2, the main benefit of the proposed method is that it processes the rejection query quickly.

### 3.3 Update Labeling Scheme and Node Relationship

In this section, update on XML documents are described by commonly used tree operations, namely, *INSERT* an element, text, or attribute, and *DELETE* an element, text, or attribute. In fact, deletion can be realized more easily. The deletion of a node will not affect the ordering of the nodes in the XML tree. However, *INSERT* operation is more complicated than *DELETE* operation.

**Definition 1.** *(Insert Operation)* For any two existing adjacent self labeling, $N_{left}$ and $N_{right}$, a new labeling $N_{new}$ can always be allocated between them without modifying already allocated labeling by the following insert operation, where len(N) is defined as the bit length.

1. Insert a node before the first sibling node, $N_{new} = N_{left}$ with the last bit “1” change to “01”
2. Insert a node after the last sibling node, $N_{new} = N_{right}$ with the last bit “1” + 1, (+ means summation, if the last bit is “9”, last bit extension “9→91”)
3. len($N_{left}$) $\geq$ len($N_{right}$) then, $N_{new} = N_{left} \oplus 1$, ($\oplus$ means concatenation)
4. len($N_{left}$) < len($N_{right}$) then, $N_{new} = N_{right}$ with the last bit “1” change to “01”

To insert a node between “1a1” and “1b1”, the size of “1a1” and “1b1”is equal, therefore we directly concatenate one more “1” after “1a1” (see Definition 1.). To insert a node between “1a1” and “1a11”, the size of “1a1” is 3 which is smaller than the size 4 of “1a11”, therefore we change the last “1” of “1a11” to “01”, the inserted label string is “1a101”. Obviously, “1a1” $\prec$ “1a101” $\prec$ “1a11” lexicographically. To insert a node between “1a11” and “1b1”, the size of “1a11” is 4 which is larger than the size 3 of “1b1”, therefore we directly concatenate one more “1a111” after “1a11”. Our proposed dynamic role based-prime number labeling scheme guarantee that we don’t have update costs in XML updating.

Using labels of the parent nodes as a part of creating labels for child nodes helps to determine the ancestor-descendant relationships and the sibling relationship between nodes. For instance, “2a1.a1”, we can understand that its parent is “1a1” and all nodes beginning with “2” are its siblings. All of its children nodes shall have “3a1a1” attached at front.

## 4 Evaluation

In this section, we compared our proposed labeling scheme in various ways. We used one PC with an Intel Pentium IV 2.66GHz CPU, 1GB Memory, and MS Windows XP. We have implemented proposed labeling scheme in JAVA (JDK1.5.0) and used SAX parser. And we used XMark [12] datasets to generate XML documents.

**Label Size and Re-Label.** We analyzed the performance of GRP [10] scheme, which supports dynamic XML data. The first ten XML documents were generated by XMark, same as that used by [10]. Then we generated labels from those documents using our proposed dynamic labeling scheme and compared with those two schemes in term of total length of labels. We discovered that our dynamic labeling scheme can be average 3 times shorter compared to GRP scheme. Detailed figures are presented in the Fig. 5. For prime number labeling are required to re-calculate is counted in Fig. 6. But our proposed labeling need not re-label the existing nodes in updates.

**Rejection Query and Query Processing Time.** Access control policy and actual number of detection of rejection queries was compared to this. The Fig. 7 demonstrates that the intended 30 rejection queries were detected 100%. Average query processing time was compared in two cases: one applied the access control method proposed in this paper and the other did not. Referring to role-based prime number which is generated before query processing, and query processing time including the pure procedure of authority checking was measured. In case of referring to access control information does not affect the system performance was discovered. Fig. 8 shows the result.

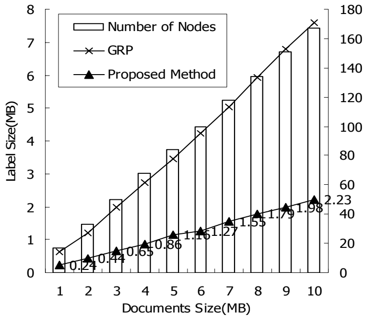


**Fig. 5.** Total Length of Label

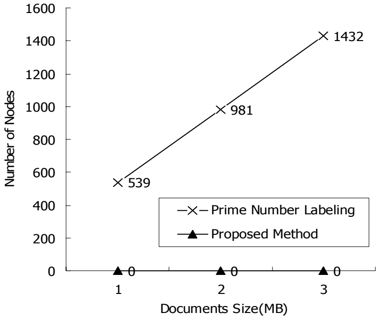


**Fig. 6.** Number of Node to Re-Label in Updates

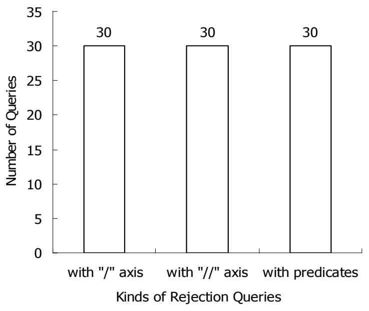


**Fig. 7.** Detection of Rejection Query

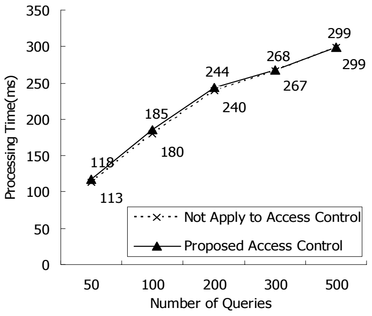


**Fig. 8.** Processing Time for Secure Query

## 5 Conclusion

In this paper, we point out the limitations of existing access control and labeling schemes for XML data assuming that documents are frequently updated. Our dynamic role-based prime number labeling scheme where labels are encoded ancestor-descendant relationships and sibling relationship between nodes but need not to be regenerated when the document is updated. Also our labeling scheme supports an infinite number of updates and guarantees the arbitrary nodes insertion at arbitrary position of the XML tree without label collisions.

In terms of security, system load is minimized and a perfect access control is implemented by application of two-step security. First of all, a query by unauthorized user is promptly rejected applying the characteristics of the prime number and a stricter access control can be applied by the application of two-step security

We will further research how to efficient secure query processing in ubiquitous sensor networks in the future.

**Acknowledgements.** This work was supported by the Korea Science and Engineering Foundation(KOSEF) grant funded by the Korea government(MOST) (No. R01-2006-000-10609-0).

## References

1. http://www.w3.org/TR/xml-fragment
2. Babcock, B., Babu, S., Datar, M., Motwani, R., Widom, J.: Models and Issues in Data Stream Systems. In: PODS (2002)
3. Bertino, E., Castano, S., Ferrari, E., Mesiti, M.: Specifying and Enforcing Access Control Policies for XML Document Sources. WWW Journal (2000)
4. Bertino, E., Ferrari, E.: Secure and Selective Dissemination of XML Documents. TISSEC 5(3) (2002)
5. Bose, S., Fegaras, L.: XFrag: A Query Processing Framework for Fragmented XML Data. Web and Databases (2005)
6. Carminati, B., Ferrari, E., Tan, K.L.: Specifying Access Control Policies on Data Streams Outsourced Data. In: Li Lee, M., Tan, K.-L., Wuwongse, V. (eds.) DASFAA 2006. LNCS, vol. 3882. Springer, Heidelberg (2006)
7. Damiani, E., Vimercati, S., et al.: Securing XML Document. In: Zaniolo, C., Grust, T., Scholl, M.H., Lockemann, P.C. (eds.) EDBT 2000. LNCS, vol. 1777. Springer, Heidelberg (2000)
8. Damiani, E., Vimercati, S., et al.: Access Control System for XML Documents. ACM Trans. Information and System Sec. 5(2) (2002)
9. Huo, H., Wang, G., Hui, X., Boning, R.Z., Xiao, C.: Efficient Query Processing for Streamed XML Fragments. In: Li Lee, M., Tan, K.-L., Wuwongse, V. (eds.) DASFAA 2006. LNCS, vol. 3882, pp. 468–482. Springer, Heidelberg (2006)
10. Lu, J., Ling, T.W.: Labeling and Querying Dynamic XML Trees. In: Yu, J.X., Lin, X., Lu, H., Zhang, Y. (eds.) APWeb 2004. LNCS, vol. 3007. Springer, Heidelberg (2004)
11. Murata, M., Tozawa, A., Kudo, M.: XML Access Control Using Static Analysis. ACM CCS, New York (2003)
12. Schmidt, A., Waas, F., Kersten, M., Carey, M.J., Manolescu, I., Busse, R.: XMark: a benchmark for XML data management. In: VLDB (2002)
13. Wu, X., Li, M., Hsu, L.: A Prime Number Labeling Scheme for Dynamic Ordered XML Trees. In: ICDE (2004)

# A Secure Mediator for Integrating Multiple Level Access Control Policies[⋆]

Isabel F. Cruz, Rigel Gjomemo, and Mirko Orsini[⋆⋆]

ADVIS Lab, Department of Computer Science, University of Illinois at Chicago
{ifc,rgjomemo,orsinim}@cs.uic.edu

**Abstract.** We present a method for mapping security levels among the components of a distributed system where data in the local sources are represented in XML. Distributed data is integrated using a semantic-based approach that maps each XML schema into an RDF schema and subsequently integrates those schemas into a global RDF schema using a global as view (GAV) approach. We transform the security levels defined on the XML schema elements of each local source into security levels on the triples of the local RDF schemas, which form a lattice. We show how the merged data in the global schema can be classified in different security classes belonging to the global partially ordered security graph.

## 1 Introduction

Data interoperation systems integrate information from different local sources to enable communication and exchange of data between them. A common model for these systems involves a global representation of the local data, which acts as a mediator for translating queries and conveying data to and from these sources using the global-as-view (GAV) approach [13]. Semantic Web languages such as RDF Schema (or RDFS) [4] and OWL [16] are particularly suited to represent the global information and to abstract from the particular data formats (relational, XML, etc.) or from the different schemas within the same format, thus addressing respectively problems of syntactic heterogeneity [6] and of structural (or schematic) heterogeneity [17].

We describe a possible scenario in which we consider two healthcare organizations, for instance a health insurance company and a hospital that want to integrate some of their patient data. The data are stored in XML. Figures 1.1 and 1.2 show respectively portions of the XML schemas of the hospital and of the health insurance company. Although data pertain to the same domain, the XML schemas display structural heterogeneities—the element *patient* is contained (nested) in the element *hospital* in one schema, while in the other schema the element *hospital* is contained (nested) in the element *customer*. In reality, the relationship between patients (or customers) and hospitals is "many-to-many" but due to the hierarchical nature of XML such relationships need to be represented using containment.

[⋆] Work partially supported by NSF Awards ITR IIS-0326284 and IIS-0513553.

[⋆⋆] Primary affiliation: Dipartimento di Ingegneria dell'Informazione, Università di Modena e Reggio Emilia, Italy. Work partially supported by MUR FIRB Networked Peers for Business project (http://www.dbgroup.unimo.it/nep4b).

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 354–362, 2008.

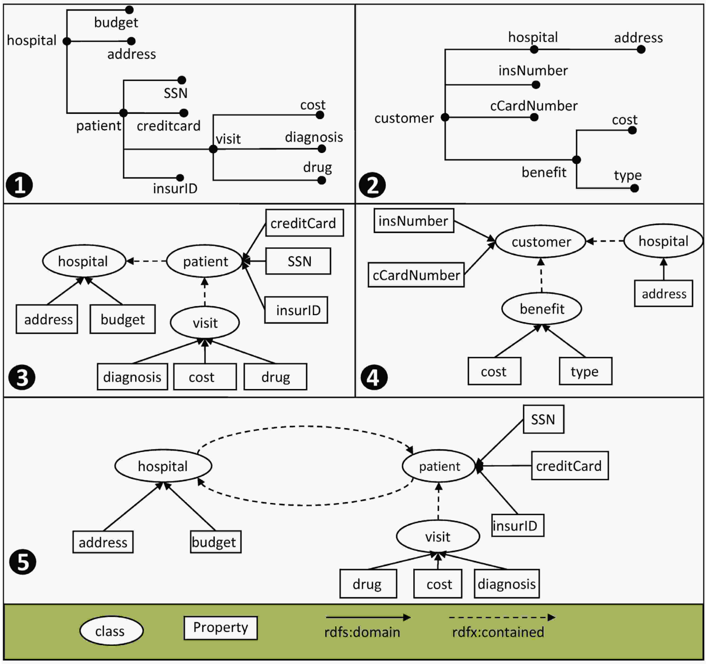


**Fig. 1.** 1. Hospital XML schema 2. Insurance XML schema 3. Hospital RDF schema 4. Insurance RDF schema 5. Global RDF schema

In addition to structural heterogeneity, our example also illustrates a case of semantic heterogeneity in that two elements that refer to the same concept have different names: *patient* and *customer*. In order to overcome syntactic, structural, and semantic heterogeneities, schemas can be integrated at a semantic level. For example, the problem of structural heterogeneities has been addressed in a previous approach [17], where a two-step integration framework is proposed. In the first step, the XML schemas are transformed into RDF schemas. RDF is a language built on top of XML, which can be used to describe relationships between entities. These relationships can be expressed in terms of *triples* of the form $(s, p, o)$. The first element, $s$, is the *subject* of the triple, the second element, $p$, is the *predicate* or *property*, and the third element, $o$, is the *object* or *value* of the property. The subject of the triple is also called *domain* of the property and the object is called *range* of the property. We define a mapping function $\mu$ next.

**Definition 1.** *The mapping function $\mu$ maps an XML schema element to an RDF schema element. If $v$ is a complex XML schema element, then $\mu(v)$ belongs to the set of RDFS classes. If $v$ is a simple XML schema element or an attribute, then $\mu(v)$ belongs to the set of RDF properties.*

As shown in Figures 1.1 and 1.3, the complex XML schema element *patient* is mapped to the RDFS class *patient*, whereas the simple XML schema element *creditcard* is mapped to the RDF property *creditCard*. As can be seen in Figures 1.3 and 1.4., the two structurally heterogeneous elements are now mapped to two different classes. A property called *rdfx:contained* is used to record the parent-child relationship between complex XML elements. The second step is that of merging the local RDF schemas into a global schema and it consists of: (1) merging of equivalent RDFS classes and RDF properties from the local sources into a single class or property on the global schema; (2) copying a class or property into the global RDF schema if an equivalent class or property does not exist. A possible global RDF schema is shown in Figure 1.5. Here the local classes *patient* and *customer* have been mapped to the global class *patient*.

The problem that we address in this paper is the security of the interoperation model described above. In particular, if the local schemas are integrated in the global schema, how can the security policy of the global schema be specified taking into account the local security policies?

We adopt a model in which each local organization enforces a multiple level access control model on its schemas [3]. In this model, data are categorized into security levels and users are assigned security clearances. We define a *partial order* or *lattice* $\preceq$ on the set of security levels as follows: given two security levels $s_i$ and $s_j$, data classified at level $s_i$ can be accessed by anyone with security clearance $s_j$, such that $s_i \preceq s_j$. The partial order can be represented by a directed acyclic graph. A chain in the graph represents a total order among the security levels along the chain.

The paper is organized as follows. Section 2 presents our security framework, including the *autonomy*, *confidentiality*, and *availability* requirements, the local security lattices and the process in which they are merged to form a global security lattice; we introduce definitions and a theorem that states that the security mappings that need to be established between two local schemas and between a local and a global schema satisfy the requirements. The last two sections, Sections 3 and 4, give a brief overview of related work, of our main contributions, and point to future work.

## 2 Security Framework

In this section we discuss the process of mapping security levels associated with the elements of the local XML schemas to the global RDF schema triples. The local security policies are represented as local security lattices associated with both the XML and the RDF schema levels. Local security lattices are merged into a global security lattice representing the global security levels associated with the global RDF schema. We assume that the only action that is permitted on the local sources is the *read* action. The results can be extended also to the *write* action, but we assume that users can only write and change the values of the local sources they are associated with. The security of the interoperation systems must satisfy the following requirements:

- *Autonomy.* The local security policies must not be affected by the security policy of the global level.
- *Confidentiality.* Given a security clearance, if a schema element is not accessible locally before the integration, then it must not be accessible after integration.
- *Availability.* Given a security clearance, if a local schema element is accessible before integration, then it must continue to be accessible after integration.

We also make the following assumptions and observations on the local XML and RDF schemas: very sensitive portions of the local XML schemas might not be shared at all; the global level contains the RDF schema, but not the instances (which reside locally). The security levels on the local XML schema elements are used to restrict access to the corresponding XML instance elements.

**Definition 2.** *A* security specification *on the XML schema tree is a pair* $[v, s]$ *where* $v$ *is a node of the local XML schema and* $s$ *is the security level associated with* $v$*. We denote the* set of security specifications *by* $S_X$.

We modify a previously proposed model to specify the security levels globally, which assigns security levels to RDFS triples based on RDF patterns [11]. Instead, we assign security levels to RDFS triples based on XML schema elements.

**Definition 3.** *A* security object *is a pair* $[t, s]$*, where* $t$ *is an RDFS triple and* $s$ *is the security level associated with* $t$*. We denote the* set of security objects of a local RDF schema *by* $S_L$.

We consider two kinds of RDF schema triples: *subject triples* and *subject-object triples*.

**Definition 4.** *A* subject triple *is an RDFS triple (s, p, o) where the subject s is a mapping* $\mu(v)$ *of an XML schema element* $v$*, and the predicate p and object o belong to the RDFS vocabulary. A* subject-object triple *is an RDFS triple (s, p, o) where the subject s and object o are two mappings* $\mu(u)$ *and* $\mu(v)$ *of two XML schema elements* $u$ *and* $v$ *which are in a parent-child or containment relationship, and the predicate p is either* rdfs:domain *or* rdfx:contained*.*

For example, *(hospital, rdf:type, rdfs:Class)* is a *subject triple* where only the subject *hospital* is mapped from an XML schema element. The triple *(creditCard, rdfs:domain, patient)* is a *subject-object triple* where the subject *creditCard* and the object *patient* are mapped from XML schema elements. Security levels assigned to *subject triples* will restrict access to information on single entities of the original XML schemas whereas in *subject-object triples* the two elements of the local XML schema may have different security levels. Accordingly, we define two security mappings that associate security specifications on the local XML schemas to security objects on the local RDF schemas.

**Definition 5.** *A* subject security mapping $\sigma$ *maps a security specification in* $S_X$ *of the form* $[v, s]$ *to a set of security objects in* $S_L$*, of the form* $[t, s]$*, such that (1)* $t$ *is a subject triple; (2)* $s$ *is the same security level for all security objects. There are, therefore, as many security objects as there are triples* $t$ *that correspond to XML schema element* $v$*. A triple* $t$ *can either correspond directly to an element* $v$ *or can be classified by inference using RDFS entailment [11].*

For instance, consider the security specification [*SSN, adm*] in Figure 2.1. The subject security mapping $\sigma$ maps that security specification to security object [*(SSN, rdf:type, rdfs:Class), adm*] in Figure 2.3 and to security object [*(SSN, rdf:type, rdfs:Resource), adm*] containing the entailed triple (because due to inheritance every *class* is also a *resource* in the RDFS model).

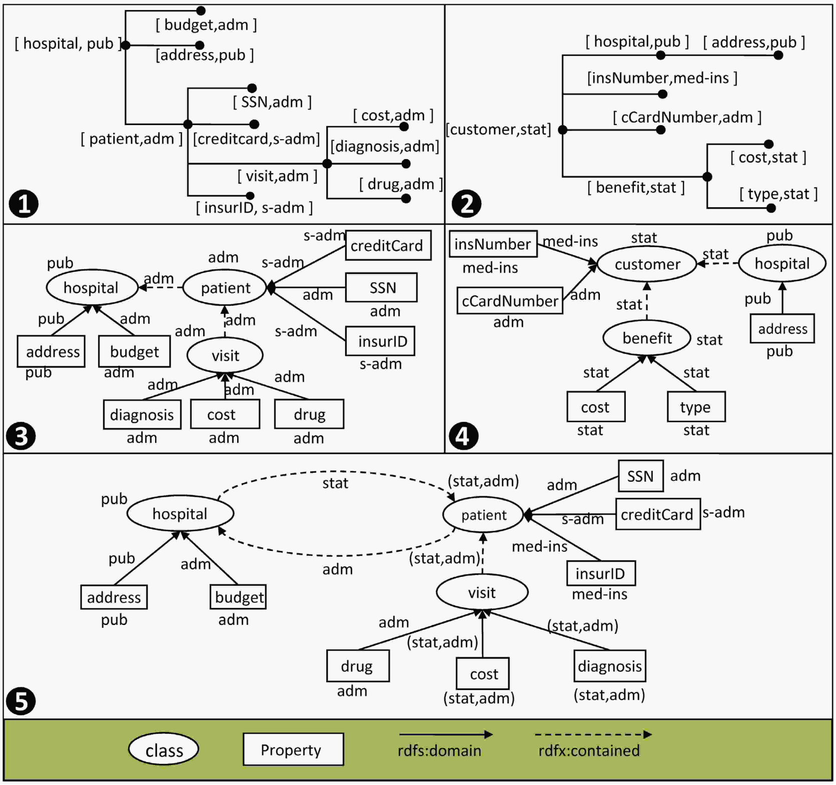


**Fig. 2.** Security levels and mappings: 1. Hospital XML schema 2. Insurance XML schema 3. Hospital RDF schema 4. Insurance RDF schema 5. Global RDF schema

**Definition 6.** *A* subject-object security mapping $\kappa$ *maps a pair of security specifications* $[v_1, s_1]$ *and* $[v_2, s_2]$ *in* $S_X$ *to a security object* $[t, s]$ *in* $S_L$*, where* $t$ *is a subject-object triple and the security level* $s$ *is the least upper bound (LUB) of the security specifications levels* $s_1$ *and* $s_2$*.*

Every *subject-object triple* is assigned to the least upper bound (LUB) of the security levels of the corresponding XML schema elements. Instead, the *subject triples* are assigned to the security level of the corresponding XML schema element. For instance, consider the security specifications [*hospital, pub*] and [*budget, adm*] in Figure 2.1. where *hospital* and *budget* are in a parent-child relationship. The subject-object security mapping $\kappa$ maps them to the security object [*(budget, rdfs:domain, hospital), adm*], if $LUB(pub, adm) = adm$. Figure 2 shows the mappings of the security specifications on the XML schemas to the security objects on the local RDFS triples.

Next, we discuss the process of merging the local security lattices into a global security lattice representing the global security levels associated with the global RDF schema, and the classification of the global RDFS triples. The merging process can be carried out by an agreement among the security administrators of the local sources.

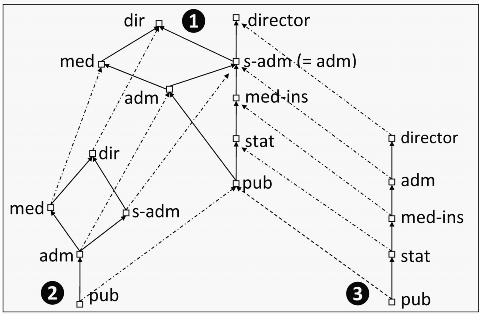


**Fig. 3.** 1. Global security lattice 2. Hospital security lattice 3. Insurance security lattice

Some local security levels from different sources may be merged in the global security lattice, while others may be just copied into it. Constraints on the orderings among security levels at the different local sources are used to define the global order. One requirement of the merging is that there are no cycles in the resulting partial order [9,14]. The partial order $\preceq$ in the local sources must also be preserved in the global security lattice. Therefore, one or more local security levels can be merged into a global security level.

**Definition 7.** *The mapping function $\theta$ maps a local security level to a global security level. The mapping function $\Theta$ maps a set of local security levels, $L_i$, to a set of global security levels $\Theta(L_i) = \{\theta(l)|\, l \in L_i\}$.*

We show an example in Figure 3 in which the dotted lines represent the mappings defined by $\theta$. The local levels *s-adm (secure administration)* and *adm (administration)* are merged into the global level *s-adm* and $\preceq$ is preserved globally. The classification of the global triples is performed by exploiting the mappings between the triples of the local and of the global RDF schemas and the mappings between the local security levels and the global ones after the merging. A global triple will be assigned a security level by taking into account the security levels of the corresponding triples in the local sources. In the most general case, the local triples mapped to the same global triple will have local security levels mapped to different global security levels. Therefore, there can be more than one candidate security level for a global triple.

**Definition 8.** *Let S be a subset of the global security levels. The source of S,* source(S)*, is the subset of S such that for each element $s_i$ in $source(S)$ there is no element $s_j$ in S such that $s_j \preceq s_i$ in the graph induced by S. Each element $s_i$ is called* minimal.

In Figure 3.1, $source(\{dir, med, \textit{med-ins}\})$ is the set $\{med, \textit{med-ins}\}$.

**Definition 9.** *Let $S_G$ be the set of security objects of the global RDF schema and $S_{Li}$ be a set of local security objects in $S_L$, where the triples in each security object in $S_{Li}$ are mapped to the same global triple $t_{g_i}$. Let $L_i$ be the set of local security levels of $S_{Li}$. The* global security mapping *$\gamma$ maps each $S_{Li}$ to a subset $S_{Gi}$ of $S_G$, whose elements share that same triple but have as security level one of the security levels in $source(S)$,* *where* $S = \Theta(L_i)$*. The cardinality of the set* $S_{Gi}$ *is the same as the cardinality of* $source(S)$*.*

For instance, consider two local triples *(cost, rdfs:domain, visit)* in Figure 2.3 and *(cost, rdfs:domain, benefit)* in Figure 2.4. that are mapped to the same global triple $t_{g1}$ = *(cost, rdfs:domain, visit)* in Figure 2.5. The global security mapping $\gamma$ maps the set $S_1$ formed by the two local security objects [*(cost, rdfs:domain, visit), adm*] and [*(cost, rdfs:domain, benefit), stat*] to the set $S_{G1}$ formed by the global security objects [*(cost, rdfs:domain, visit), stat*] and [*(cost, rdfs:domain, visit), adm*], because $source(S_1) = \{adm, stat\}$.

**Theorem.** *Assuming security autonomy after source integration, the local security mappings* $\sigma$ *and* $\kappa$ *and the global security mapping* $\gamma$ *preserve data confidentiality and availability.*

**Proof Sketch.** By means of the local security mappings $\sigma$ and $\kappa$, the local security levels are mapped either to themselves (in the case of a subject triple) or to their least upper bound (in the case of a subject-object triple). Given two local security levels $l_1$ and $l_2$, we have $l_1, l_2 \preceq LUB(l_1, l_2)$. The global security mapping $\gamma$ maps a set of local security objects to a set of global security objects where the global security levels are minimal. It may be that a global triple is associated with a global security level $g \preceq LUB(l_1, l_2)$, but due to the security autonomy of the local sources the local triple will remain classified at level $LUB(l_1, l_2)$. Therefore, if an XML schema element cannot be accessed before integration, it will continue to be inaccessible afterwards, thus guaranteeing the confidentiality of the data.

Through the subject security mapping $\sigma$, the local security level remains the same, therefore the XML schema element remains available. Subject-object security mapping $\kappa$ maps two security specifications to a security object, therefore the security level obtained may be more restrictive. This type of mapping deals with the security of the relationship between the subject and the object elements. Even if the relationship is restricted, they can always be accessed individually at the corresponding single triples' security levels. The global security levels obtained by the global security mapping $\gamma$ are minimal because some local triples are classified at those minimal security levels. Therefore, the minimal global security levels guarantee the availability of the data.

## 3 Related Work

**XML Access Control Models.** XML access control models have been the focus of recent research, including approaches in which the access control model is expressed in terms of tuples that specify who can access which schema element, what type of access is allowed, and how the access rights propagate on the XML tree [2,7,8].

**RDF/S Access Control Models.** A method for transforming RDF graphs into trees so as to hide subtrees of a given node has been proposed [12]. Related work includes the work by Farkas and Jain [11] that has been mentioned in Section 2.

**Secure Interoperation Models.** The approach by Pan *et al.* uses a mediator among database systems in an RBAC access control model and mappings between roles in different local sources [15]. In another approach, Candan *et al.* propose a secure interoperation model where a global mediator can enforce global access control rules, or just be a conveyer of the information exchanged between the local sources [5]. Bonatti *et al.* propose the merging of sets of ordered security levels using a logic programming approach [3]. In other work, Dawson *et al.* propose a framework for secure interoperation between local applications mediated by a global application [9]. The work of Farkas *et al.* is the closest to ours [10]. However, they use a "top-down" approach in which they start from the RDF global schema, whereas we start from the XML sources. Another difference is that they use discretionary access rights, whereas we use multiple level security lattices.

## 4 Conclusions and Future Work

We have proposed a translation model for security levels from local XML schema sources to a global RDF schema. We follow a bottom-up approach and respect the principle of local autonomy in that local security policies continue to be valid. In the future, we will consider the implications of having specifications of security levels not only on the XML schema elements, but also on their instances. We will expand our approach to full XML schemas, including for example IDREF tags. We will also investigate how this approach can be generalized to other data representation models. Furthermore, we plan to incorporate our model into the MOMIS system [1].

## References

1. Beneventano, D., Bergamaschi, S., Vincini, M., Orsini, M., Mbinkeu, R.C.N.: Getting through the THALIA benchmark with MOMIS. In: International Workshop on Database Interoperability (InterDB) co-located with VLDB (2007)
2. Bertino, E., Castano, S., Ferrari, E., Mesiti, M.: Protection and administration of XML data sources. Data and Knowledge Engineering 43(3), 237–260 (2002)
3. Bonatti, P.A., Sapino, M.L., Subrahmanian, V.S.: Merging heterogeneous security orderings. Journal of Computer Security 5(1), 3–29 (1997)
4. Brickley, D., Guha, R.: RDF Vocabulary Description Language 1.0: RDF Schema. W3C Working Draft (February 2004), http://www.w3.org/TR/rdf-schema
5. Candan, K.S., Jajodia, S., Subrahmanian, V.S.: Secure mediated databases. In: IEEE International Conference on Data Engineering (ICDE), pp. 28–37 (1996)
6. Cruz, I.F., Xiao, H.: Using a Layered Approach for Interoperability on the Semantic Web. In: Int. Conf. Web Information Systems Engineering (WISE), pp. 221–232 (2003)
7. Damiani, E., De Capitani di Vimercati, S., Paraboschi, S., Samarati, P.: A fine-grained access control system for XML documents. ACM Trans. on Information and System Security 5(2), 169–202 (2002)
8. Damiani, E., Samarati, P., De Capitani di Vimercati, S., Paraboschi, S.: Controlling access to XML documents. IEEE Internet Computing 5(6), 18–28 (2001)
9. Dawson, S., Qian, S., Samarati, P.: Providing security and interoperation of heterogeneous systems. Distributed and Parallel Databases 8(1), 119–145 (2000)
10. Farkas, C., Jain, A., Wijesekera, D., Singhal, A., Thuraisingham, B.: Semantic-aware data protection in web services. In: IEEE Workshop on Web Service Security (2006)

11. Jain, A., Farkas, C.: Secure resource description framework: an access control model. In: ACM Symp. on Access Control Models and Technologies (SACMAT), pp. 121–129 (2006)
12. Kaushik, S., Wijesekera, D., Ammann, P.: Policy-based dissemination of partial web-ontologies. In: Workshop on Secure Web Services (SWS), pp. 43–52 (2005)
13. Lenzerini, M.: Data integration: a theoretical perspective. In: ACM Sigact-Sigmod-Sigart Symp. on Principles of Database Systems (PODS), pp. 233–246 (2002)
14. Oliva, M., Saltor, F.: Integrating security policies in federated database systems. In: Annual Working Conf. on Database Security (DBSec), pp. 135–148 (2000)
15. Pan, C.-C., Mitra, P., Liu, P.: Semantic access control for information interoperation. In: ACM Symp. on Access Control Models and Technologies (SACMAT), pp. 237–246 (2006)
16. Smith, M.K., Welty, C., McGuinness, D.L.: OWL web ontology language guide (February 2004), http://www.w3.org/TR/owl-guide/
17. Xiao, H., Cruz, I.F.: Integrating and exchanging XML data using ontologies. In: Spaccapietra, S., Aberer, K., Cudré-Mauroux, P. (eds.) Journal on Data Semantics VI. LNCS, vol. 4090, pp. 67–89. Springer, Heidelberg (2006)

# Domain Account Model

Giulio Galiero, Paolo Roccetti, and Andrea Turli

Engineering Ingegneria Informatica SpA, Roma RM 00185, Italy
{giulio.galiero,paolo.roccetti,andrea.turli}@eng.it

**Abstract.** The use of gateways as a user-friendly way to access the Grid is increasing, as evidenced, for example, by the popularity of TeraGrid Science Gateways. Such gateways, however, imply additional layers of software abstraction, which in turn implies more levels of trust delegation - thus compounding security problems of enforcing trust at different layers.

In this paper we present the Domain Account Model (DAM), extending the Shibboleth and GridShib ones to enable interoperability between identity-based (i.e. GSI) and attribute-based (i.e. SAML) Grid authorization mechanisms, to ease the administration of user attributes by allowing domain separation between Real and Virtual Organizations (VO), and to improve the trust management by means of the OAuth protocol.

## 1 Introduction

Identity fragmentation over different administrative domains has recently been recognised as one of the main usability issues of multidomain systems [1]. The lack of consistent identity management in Grids denies users a seamless experience in the access and usage of resources belonging to different administrative domains. This problem has been addressed in grid systems [2] by leveraging Public Key Infrastructures (PKI) to provide each user with credentials that can be shared among different domains. A limitation of this solution, however, is the lack of privacy that comes from the need to expose the entire user certificate to other parties.

In addition, users must manage their own credentials, which limits the wide usage of grids. A common solution, for example, is to store credentials in an online repository (e.g. MyProxy [3]) and delegate them to portals and services. The underlying security mechanisms in such systems, however, are not transparent to end users, and often require significant user effort to operate.

This paper first enlarges upon existing approaches to these problems (section 2), highlighting additional open issues, and then proposes the Domain Account Model (DAM) as a potential solution (section 3).

## 2 Existing Models

In recent years various models have been put forward to handle Identity Management issues. In Shibboleth [4] and OpenID [5], Identity Providers (IdP) are responsible for authenticating users within their home domain and releasing signed

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 363–370, 2008.

SAML assertions of authenticity (Security Assertions Markup Language [6]). These assertions can be used by federated resources, namely Service Providers (SP), to enforce authorization at resource domains. Shibboleth also allows users to control which attributes can be released to each SP, thus preventing unneeded disclosure of user information.

The Community Account Model (CAM) [7] (developed as part of the TeraGrid project) employs the notion of a web gateway as a usable and scalable solution to access grid resources. Usability is achieved by authenticating users into the TeraGrid 'Science Gateway' with traditional methods (e.g. username/password) - so that users no longer need to possess and manage an X.509 certificate as an explicit authentication/autorization token ("X.509 unawareness"). Scalability is obtained through attribute-based authorization. On the down side, CAM only allows coarse grained authorization to be enforced on grid services, as all community users access protected resources using the identity of the Gateway (whose credentials are located at the gateway itself).

The limitations of CAM are addressed in the GridShib project [8] which integrates the gateway approach with the IdP paradigm provided by the Shibboleth architecture [4]. SAML assertions released by the IdP are pushed to the Shib-enabled Gateway for authorization. Once assertions are verified by the SP, the GridShib SAML tool (GS-ST) enables the Gateway to bind them to new proxy credentials. These credentials can then be used to authenticate to a Shib-enabled Resource running in a Globus Toolkit (GT) container. When a request is received by the container, credentials are parsed by the GridShib For Globus Toolkit (GS4GT) plugin to extract IdP assertions. This allows resource providers to enforce authorization on a user-attribute basis [9]. The user can also create SAML attributed proxy credentials on-demand using the Shib-enabled GridShib Certification Authority (GS-CA). This component authorizes users according to SAML assertions released by the IdP, and embeds these assertions in short-lived proxy credentials. The user can then contact resources directly, i.e. bypassing the Science Gateway.

### 2.1 Open Issues

The identity management mechanisms described in the previous section protect user privacy and increase the usability of grid resources. Nevertheless, our own experience with these models has revealed some open issues.

First of all the model depicted by GridShib only refers to containers capable of handling user attributes as SAML assertions; therefore it cannot be considered as a global comprehensive solution. Typically, resources accessible through a Gateway are deployed in containers belonging to different administrative domains. Each domain can apply local policies for managing and upgrading its own containers. Thus, it is likely that some of these resources will still be protected using identity-based authorization (e.g. gridmap-files), which is not well supported by GridShib. In fact, within the GridShib model resources can be accessed in two ways: either (1) through the Gateway, using proxy credentials of the Gateway itself, or (2) directly, using proxy credentials released by the GS-CA. As such, there is an inconsistency in accessing gridmap-protected resources.

A second limitation of the Shibboleth and Gridshib models is the poor support for Virtual Organizations (VOs). VOs are a foundational concept in grid computing: they are made up of different Real Organizations (ROs) members, entitled to access and share resources. VO-specific attributes are usually assigned to VO members (e.g. roles, as in the RBAC model [10]) to enable a scalable VO administration. In existing models, VO-specific attributes need to be maintained and released by the IdPs of ROs. However this solution faces scalability problems as the number of VOs and of ROs involved in the VO increases - since each SP needs to trust every IdP of every RO involved in the VO. The presence of multiple, distributed IdPs likewise compounds the problem of VO administration.

Finally, existing models do not offer a solution to limit the actions performed by entities delegated by the users. Access to the basic Grid services is typically achieved by stacking up software layers in increasing levels of abstraction. This enhances usability by hiding the innermost complexities of the infrastructure, but also increases the levels of trust delegation: a multi-layered architecture needs to implement security mechanisms to enforce trust at each level. In certain scenarios, user's consent must be requested when a delegated entity acts on her behalf in order to prevent malicious behaviour and increase trustworthiness.

## 3 The Domain Account Model

In this section we introduce the Domain Account Model (DAM) - which builds upon existing models to address the open issues described in 2.1.

The main innovation in the DAM is the adoption of Virtual Organizations as contexts of resource sharing. VO resources can be accessed by means of VOs easy-to-use graphical interface, known as gateways. Usability can greatly benefit from gateways, but this comes at a cost to security - since the gateway is enabled to act on the user's behalf, once she has logged into the gateway. In DAM, however, the Gateway has less autonomy and acts more as an intermediary, enabling users to access resources in the context of a VO. It is worth noting that a single Gateway can be bound to a number of VOs. The adoption of VOs in this way, has three main consequences.

Firstly, there is the need to support interoperability between identity-based and attribute-based authorization. In DAM, user proxy credentials released by a VO-specific Certification Authority (VO-CA) are required to access resources (either directly by users, or by the gateway on the users behalf).

Secondly, there is the need to separate user attributes into two sets: RO-specific attributes and VO-specific ones. RO-specific attributes are maintained and released by the IdP of the RO the user belongs to, while VO-specific attributes are released by an Authorization Authority managed in the context of the VO (VO AA). To access VO resources the Gateway needs to include VO-specific attributes in the users proxy credentials.

Thirdly, as the Gateway is just an intermediary between users and resources, there is a need to support advanced trust scenarios between the user and the Gateway itself. The approach taken in DAM entails controlling the gateway access both to the VO CA and to the VO AA. Two main scenarios have been identified, depending on the level of trust the user places on the gateway. In the first scenario (*user-gateway full trust*) users would completely trust the gateway once logged in - that is; the mere act of logging into the portal is an explicit delegation of complete trust to the gateway. If total trust is enforced, the gateway can retrieve any user VO-specific attribute from the VO AA, without asking for user approval. In the second scenario (*user-gateway partial trust*) users would not entirely trust the gateway. The user is able to control the activities the gateway is performing on her behalf. In order to approve, restrict or deny the gateway request, however, the user must be informed of which attributes the gateway is aking to retrieve. This raises issues of privacy relating to the user attributes stored in the VO AA. As described shortly, OAuth [11] specifications offer one solution to this - in brief; third-parties, acting on behalf of the user, can retrieve sensitive data given explicit approval from the user.

A general diagram of the DAM is shown in Figure 1. Subsequent sections describe in more detail how DAM works.

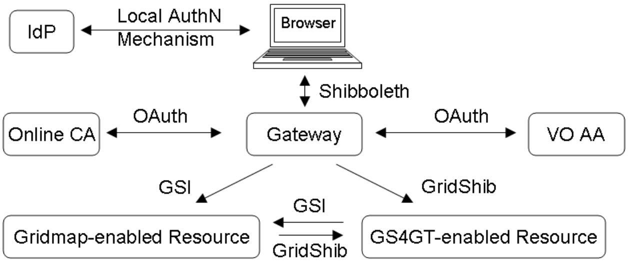


**Fig. 1.** DAM Overview

### 3.1 Identity- and Attribute-Based Authorization Interoperability

A key advantage of the DAM proposal is that, by delegating user credentials to the Gateway, the Gateway is able to access resources protected with identity-based mechanisms. Thus DAM supports interoperability by obviating any need for the Gateway to use its own identity to access resources. The delegation process is shown in Figure 2.

The user authenticates herself to the IdP and obtains SAML assertions she can send to the Gateway for authorization (Steps 1 and 2). After authorization, the Gateway asks the Online CA (could be a GridShib-CA instance) to issue new user proxy credentials. This request to the Online CA also includes SAML assertions received from the user in Step 2. The Online CA replies with new proxy credentials containing (1) the user's Distinguished Name (DN) as subject, and (2) SAML assertions received from the Gateway as extensions (Step 3). The Gateway uses these proxy credentials for resource access authentication. If the resource is protected with an attribute-based mechanism (e.g. GridShib GS4GT handler), then authorization is based on user attributes contained in

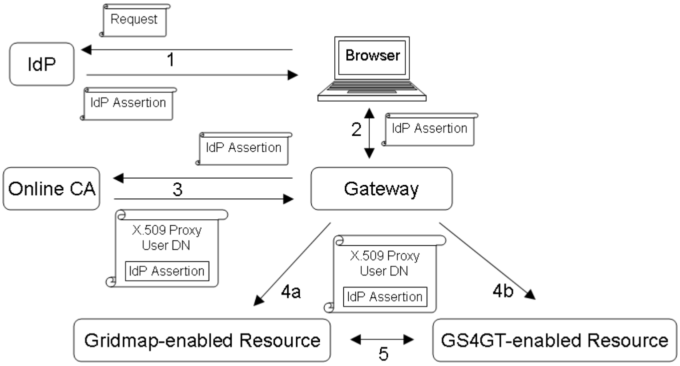


**Fig. 2.** Identity- and Attribute-based Authorization Interoperability

a certificate extension (Step 4a). If the resource is instead protected with an identity-based mechanism, (e.g. GSI GridMap file) then authorization is based on the certificate subject (Step 4b). The Gateway can also delegate credentials to resources (e.g. by using GSI delegation), thus enabling interoperability between resources protected with identity-based and attribute-based mechanisms (Step 5).

In Figure 2 the *user-gateway full trust* scenario is assumed in the interaction between the Gateway and the VO CA. If the gateway is not fully trusted by the user, as in the *user-gateway partial trust* scenario, then the interaction between the Gateway and the VO CA can be subject to user's consent, as will be explained shortly (section 3.3).

### 3.2 Real and Virtual Organization Domains

Separation of administrative domains is a desirable feature in order to achieve flexibility and efficiency in attributes administration. DAM extends the federation model by adding an extra layer supporting management of the user's VO-specific attributes. Each different federated administrative domain can manage RO-specific user attributes (i.e. in a LDAP server) whereas VO-specific user attributes are kept in a separate repository belonging to the VO domain. In this way, a VO domain can define its own policies for controlling resource access (e.g. by using the eXtensible Access Control Markup Language, XACML [12]).

In DAM, each administrative domain maintains full control over internal user privileges by defining RO-specific user attributes in IdP. VO-specific user attributes are defined in the VO Attribute Authority within the VO domain. By introducing this jurisdiction separation, VO resources need only trust the VO AA, instead of all real organization IdPs.

As in CAM, the DAM approach adopts the use of Shib-enabled Gateways for transparent access to different VO resources, but enriches this use with the domain separation model - as depicted in Figure 3 and explained below.

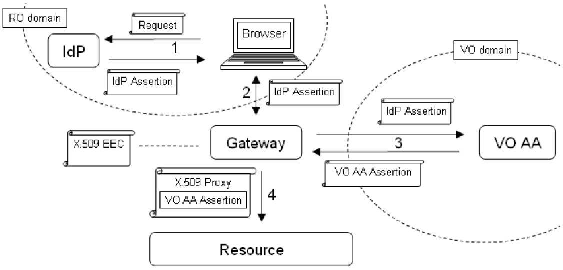


**Fig. 3.** Separation between RO and VO domains

During the authentication step against her IdP (step 1) the user retrieves IdP attributes as SAML assertions. The user then delegates these IdP attributes to the gateway (step 2), at which point the gateway can retrieve VO-specific user attributes from VO AA (step 3). Finally (step 4), the gateway binds the VO-specific user attributes to the gateway proxy certificate (alternatively, a user proxy certificate may be obtained from the VO CA, as depicted in Figure 2).

This approach has two key advantages: (1) the VO authorization mechanism is based on VO domain policies, thus VO resources can enforce authorization locally [13]; and (2) domains remain protected even if an IdP is compromised.

In Figure 3 the *user-gateway full trust* scenario is assumed in the interaction between the Gateway and the VO AA. If the gateway is not fully trusted by the user, as in the *user-gateway partial trust* scenario, then the interaction between the Gateway and the VO AA can be subject to user's consent, as will be explained shortly (section 3.3).

### 3.3 User-Gateway Partial Trust Scenario in DAM

In DAM the RO and the VO domains are linked together by a Gateway which acts as a bridge. In order to meet the requirements addressed by the *user-gateway partial trust* scenario, the DAM also integrates the OAuth model. OAuth is an open protocol defining how a user can authorize a consumer to access the user's resources protected by a service provider (SP), without requiring the user to disclose her own SP credentials to the consumer. Thus, OAuth can act as a means to enforce user control interaction upon VO-specific attributes retrieval. In DAM terms, the consumer is embodied in the gateway, while the service provider is the VO AA (storing the user's attributes as protected resources).

As described in Figure 4 the user first logs into the gateway (step 1). The gateway then requests user attributes from the VO AA (step 2) - which returns a Request token as response. This token is redirected to the user for approval (step 3). Provided that the VO AA is Shib-protected, the user can authenticate locally at her own IdP (step 4) and then present the IdP authentication assertion at the VO AA along with the request token approval/denial. If approved, the VO AA exchanges the request token with an access token and passes it to the gateway. The access token is eventually used by the gateway to actually retrieve the VO-specific user's attributes.

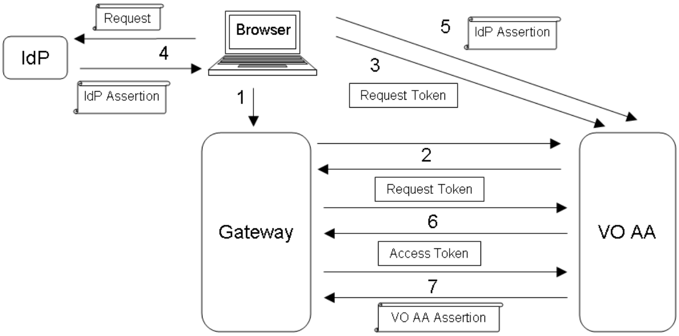


**Fig. 4.** OAuth model integration in DAM

The *user-gateway partial trust* scenario can also act as a means to enforce user control interaction upon user's proxy credentials retrieval. In this case the interaction flow is the same described in Figure 4, but the service provider is now embodied in the VO CA.

## 4 Conclusions

In this paper we introduced the Domain Account Model (DAM) as an extension of existing CAM and GridShib models. The motivations behind the DAM design are threefold. First, to support interoperability between attribute and identity based authorizations. Second, to enable and contextualize resource access for users belonging to different administrative domains. Third, to improve the user's control over resource usage by delegated entities.

An outstanding non-technical issue concerns the integration of DAM with grids regulated by policies defined by the International Grid Trust Federation (IGTF) [14] (such as EGEE [15] and OSG [16]). The DAM's use of online Certificaton Authorities does not comply with such regulations. Future work will focus on experimenting DAM within different scenarios in which VOMS [17] serve as the VO AA. Planned extensions include the definition of a VO policy framework based on a standard language (e.g. XACML), and the addition of auditing functionalities.

## Acknowledgements

This work is carried out by the Grid Research Unit of Engineering Ingegneria Informatica S.p.A. (http://www.eng.it) in the context of the GriFin - Grid For Finance project (http://www.grifin.eu).

"Globus Toolkit" is a registered trademark of the University of Chicago.

"Shibboleth" is a registered trademark of Internet2.

## References

1. Introduction to the laws of identity (2005), http://www.identityblog.com/stories/2005/05/13/TheLawsOfIdentity.pdf
2. Foster, I., Kesselman, C., Tuecke, S.: The anatomy of the Grid. International J. Supercomputer Applications 15 (2001)
3. MyProxy Credentials Management Service, http://grid.ncsa.uiuc.edu/myproxy
4. Shibboleth Architecture, http://shibboleth.internet2.edu/docs/internet2-mace-shibboleth-arch-protocols-200509.pdf
5. OpenID, http://openid.net
6. OASIS Security Services (SAML) Technical Committee, http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=security
7. Welch, V., Barlow, J., Basney, J., Marcusiu, D., Wilkins-Diehr, N.: A AAAA model to support science gateways with community accounts. Concurrency and Computation: Practice and Experience 19(6), 893–904 (2006)
8. GridShib Deployement Scenarios, http://gridshib.globus.org/about.html#gridshib-deploy
9. Scavo, T., Welch, V.: A Grid Authorization Model for Science Gateways. Concurrency and Computation: Practice and Experience (to appear), http://gridfarm007.ucs.indiana.edu/gce07/images/e/e4/Scavo.pdf
10. Ferraiolo, D., Kuhn, R.: Role-based Access Control. In: Proceedings of 15th National Computer Security Conference (1992)
11. OAuth Core 1.0 Final Specfications, http://oauth.net/core/1.0/
12. XACML 2.0 Core: Specification Document, http://docs.oasis-open.org/xacml/2.0/access_control-xacml-2.0-core-spec-os.pdfvo
13. Core and hierarchical role based access control (RBAC) profile of XACML v2.0, http://docs.oasis-open.org/xacml/2.0/access_control-xacml-2.0-rbac-profile1-spec-os.pdf
14. International Grid Trust Federation, www.gridpma.org/IGTF-Federation-Constitution.pdf
15. Enabling Grids for E-science (EGEE), http://www.eu-egee.org/
16. Open Science Grid, http://www.opensciencegrid.org
17. Alfieri, R., Cecchini, R., Ciaschini, V., dell'Agnello, L., Frohner, Á., Gianoli, A., Lörentey, K., Spataro, F.: VOMS, an Authorization System for Virtual Organizations. In: Fernández Rivera, F., Bubak, M., Gómez Tato, A., Doallo, R. (eds.) Across Grids 2003. LNCS, vol. 2970, pp. 33–40. Springer, Heidelberg (2004)

# Impostor Users Discovery Using a Multimodal Biometric Continuous Authentication Fuzzy System

Antonia Azzini and Stefania Marrara

Department of Information Technology, University of Milan,
via Bramante 65, 26013 Crema (CR), Italy
{azzini,marrara}@dti.unimi.it

**Abstract.** In the last few years the security of the user's identity has become of paramount importance. In this paper we investigate the behavior of a fuzzy controller with a multimodal biometric system as input designed with the aim of preventing user substitution after the initial authentication process. In particular this paper presents the results of the system behavior tested with impostor users.

## 1 Introduction

Recently, access control has become one of the most important issues in application design. In many contexts it is of paramount importance to be sure that a person is who he/she claims to be, therefore access control authentication plays a main role. In literature [15] authentication techniques are mainly divided into two main classes: *weak* and *strong* authentication methods. Weak authentication is based on knowledge, such as a password, or on a token, e.g., a key, magnetic or chip card. In recent years, these traditional methods for authentication have been shown to suffer flaws in security. Such flaws include forgotten or easily guessed passwords, PIN numbers written on the back of cards, etc. Therefore, there have been an intense study of alternative methods to alleviate problems.

Strong authentication methods have been developed to address drawbacks of traditional techniques [13]. They include biometric systems, such as physical biometrics (fingerprints [10], hand or palm geometry and retina, iris or facial characteristics [6]) or behavioral features (signature, voice, keystroke pattern and gait). Biometrics is based on the fact that a person possesses certain characteristics — such as retinal patterns, fingerprint patterns, gait, etc. — that are biologically or behaviorally unique to an individual. The user's claimed identity is corroborated by these characteristics, rather than a forgettable password.

In highly sensitive environments, a single, initial authentication may not be sufficient to guarantee security, but it may be necessary to check the identity of the session working user many times (e.g., at random intervals) to prevent identity substitution after the initial authentication step.

Our research [1,2,3] proposes a methodology, based on biometric authentication techniques, studied to guarantee a high and prolonged-in-time level of

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 371–378, 2008.

security. Focus of this paper is the study of the behavior of the multimodal system proposed in [3] in case of malicious users who try to substitute the authenticated ones. In such a scenario, the system must distinguish between the initial authentication phase, in which it recognizes the user profile and allows the access, and the subsequent authentication steps, in which the system continuously checks the trustworthiness level and decides if its trust in user's identity is high enough to allow the user to continue performing the current activity.

The structure of the paper is as follows. Section 2 describes the general architecture of the fuzzy methodology used to ensure the user identity during time, Section 3 shows some experimental results used to test the behavior of the system when a malicious user tries to substitute the authenticated one and, finally, Section 4 reviews the conclusions of this work and proposes some future work and open issues.

## 2 A Fuzzy Controller for Continuous Multi-modal Authentication

In this section we briefly overview the multimodal system proposed in [3] that has been used to perform the study target of this work. In [3] a fuzzy controller that computes an output variable expressing trust in the user identification was presented. The proposed biometric system exploits two different biometric devices (a face recognizing device and a fingerprint based device) achieving an effective continuous biometric authentication ruled by a fuzzy controller. The face trait is adopted in order to ensure a continuous control of the face of the users. The fingerprint system can be used in a non synchronously manner when specific conditions occur.

In order to improve the privacy of a user during the overall system, the biometric authentication process is carried out at local level, by using, if needed, on-board biometric authentication devices. In this case the identification data that are sent to the fuzzy system represent the result of the biometric matcher, and correspond to the trustworthiness evaluation parameters that will be processed by the fuzzy controller in order to obtain the trust value.

Our controller follows the Mamdani approach with a *center of area* defuzzification [11,5]. Indeed, when applied to trust-based decision support systems, Mamdani is an intuitive approach since the output is modeled using linguistic variables rather than linear or quadratic equations. Experimental results confirm the goodness of our choice (see [3]).

A detailed description of the biometric multi-modal system considered in this work has been presented in [4]. We suppose that, after the initial authentication in which the user enters a password useful to individuate the templates for the biometric matching phases, the system checks the identity of the user on the basis of the only face recognition matching value *BIOFACE*. If *BIOFACE* goes down a certain threshold, then the fuzzy controller intervenes and a new fingerprint acquisition is required. At this point the fuzzy controller computes its trustworthiness in the user identity on the basis of face and fingerprint matching scores. If the trust value is really too low then the session ends, otherwise the control can go back to the fuzzy controller for a new trust value computed using new acquisitions for face and fingerprint, or back to the face recognition device. In the latter case the face recognition system continually provides new biometric matching values till *BIOFACE* goes again down the threshold and the cycle restarts or the session time ends. This cycle ends when the session is closed by the system, or ended by the user.

## 2.1 Architecture of the Fuzzy Controller

The core of the approach is a trust evaluation process that continuously checks the identity of the user who is performing a certain activity. After receiving the initial authentication, the server accepts or refuses the user on the basis of the biometric matching values (*BIOFACE* and *BIOFINGER*) computed by the multimodal biometric system in input. *BIOFACE* has to be higher than a certain threshold (*th*), which is fixed for the application. In case the user is authenticated, the system receives one value *BIOFACE*, defined in the range of the possible values assumed by the fuzzy membership functions $[0, 100]$, (e.g., 85) which represents how the biometric acquisition matches the user's template. Then, the user identity is kept under control by a face recognition system that continuously proposes new values for *BIOFACE* during the entire working session. If, at time *t*, *BIOFACE* goes under a certain threshold *th*, the system requires two new acquisitions for *BIOFACE* and *BIOFINGER* and the fuzzy controller is activated to compute a *TRUST* value. The *TRUST* value is the parameter used by the system to choose among three different options:

1. end of session (*TRUST* in user's identity is too low);
2. new fingerprint and face acquisition (*TRUST* is low and needs to be confirmed)
3. new face acquisition (*TRUST* is enough high to trust only the face recognition system to ensure the user identity)

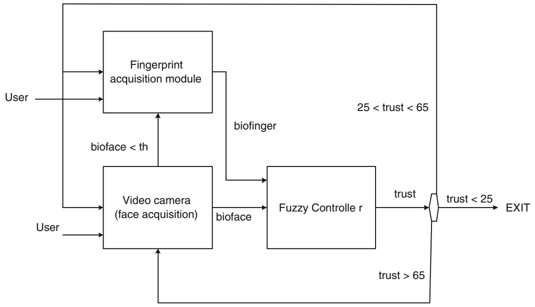


**Fig. 1.** Architecture of the Multi-modal Fuzzy Trust Model

### 2.2 Fuzzy Controller Operations

The entire process implemented in our approach is shown in Figure 1. At the core of the system, the information obtained by the biometric engines at local level, i.e. the value *BIOFACE* and *BIOFINGER*, are fed into a fuzzy inference engine in order to calculate a trust value *TRUST* that expresses the level of trust of the system in the user's identity after the initial authentication at time $t_0$, when *BIOFACE* and *BIOFINGER* are initialized; at each time $t_i$ $(i > 0)$, a new value of *BIOFACE* keeps under control the user identity and in case it goes down a threshold *th* the system asks for a fingerprint acquisition. In this case a *TRUST* value is computed by the fuzzy inference engine, and then defuzzified through a *defuzzifier* engine, using the standard centroid-of-area technique. The output is then fed to a decision point to compute the next step at time $t_{i+1}$. If the output trust is *low* the system asks for trust enforcement by going through the acquisition phase. In this case the system asks for a user re-authentication, that can be face, fingerprint-based or based on both, depending on the *TRUST* defuzzified value at that time $t_{i+1}$. In particular, the system re-acquires the parameters *BIOFACE* and *BIOFINGER* if the *TRUST* value at time $t_{i+1}$ belongs to a certain range (e.g., $[25, 65]$ are the values experimentally chosen for our prototype, see Section 3) previously defined, while re-acquires only *BIOFACE* if the *TRUST* value at time $t_{i+1}$ is higher than the range high threshold. When the trust level decays to the value of *very low*, or when the maximum value of the examination time is reached, the execution step goes to the end of the session and the process stops.

## 3 Experiments

The simulations of the multi-modal fuzzy trust controller have been carried out by considering the cases of genuinge-impostor changes, after a first good authentication with a genuine user. In this approach the datasets used respectively for face and finger authentication show the behavior of different impostors that try to access the system during a genuine working session. In particular, the simulation of the matching score of a fingerprint biometric system can be effectively achieved under the following hypothesis:

- the different pressures of the user fingertip on the sensor are independent among them,
- both 'genuine' or 'impostor' match scores follow a normal distribution.

Without lack of generality, in the followings, means and standard deviations for 'genuines' and 'impostors' are estimated on data obtained from real biometric systems e.g., [14,9,8,7]. Figure 2 shows the distributions obtained by simulating the match score of 360,000 fingers (50% genuines, 50% impostors).

The match scores of a biometric system for face authentication can be effectively simulated once the experimental conditions are fixed. In our work we considered a face biometric system authenticating a user sit in front of his/her console. Without

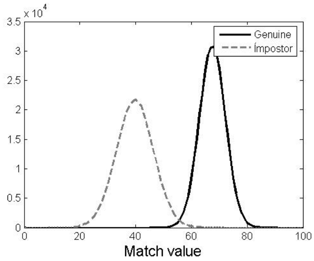


**Fig. 2.** The distribution of Genuine and Impostor of the simulated fingerprint biometric system

lack of generality, we assumed that the biometric system uses a camera placed in front of the user (e.g., a webcam on the top of the console's monitor). We assume that the same camera has been used to enroll the user, at least the first time the user has been checked by the fingerprint biometric system.

When the user is sit in front of the console, the face biometric system processes frames that are very similar to the ones used during the enrollment phase (assuming very low variability of the face during the day, no particular facial expression of the user and no evident ambient/light variations). In this case, the matching values tend to be close to the 100% perfect score. If the activities of the user requires some rotations of the head, or movements from the desk, the matching score suddenly degrades. We assumed three different user states:

- working in front of the console,
- making a phone call,
- moving away from the desk/camera.

The transitions between the three states occur with a fixed probability (let assuming for example that moving out from the office is less likely than placing a phone call and that working in front of the console is more likely than making a phone call). As in the case of the fingerprint simulation, the face matching score of each state is assumed to be distributed as a normal. Figure 3 contains the obtained matching score of a section of a genuine-impostor simulation, by showing the decrease of matching values when a malicious user attempts to corrupt the system (Figure 3(b)) respect to the normal activity carried out by an authorized user (Figure 3(a)).

### 3.1 Discussion and Results

A real-world case in which permissions and identity of one hundred of users have to be checked during one hour working sessions is considered. The aim of this approach is to ensure a continuous check of the user identity, without interrupting the user working-session with a high number of finger authentication requests, in the case of right user, or by stopping the entire working session if an attempt of intrusion occurs.

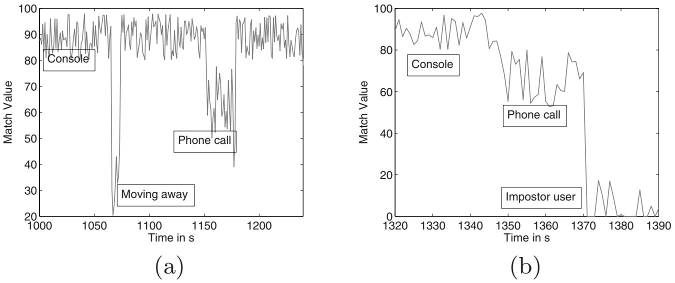


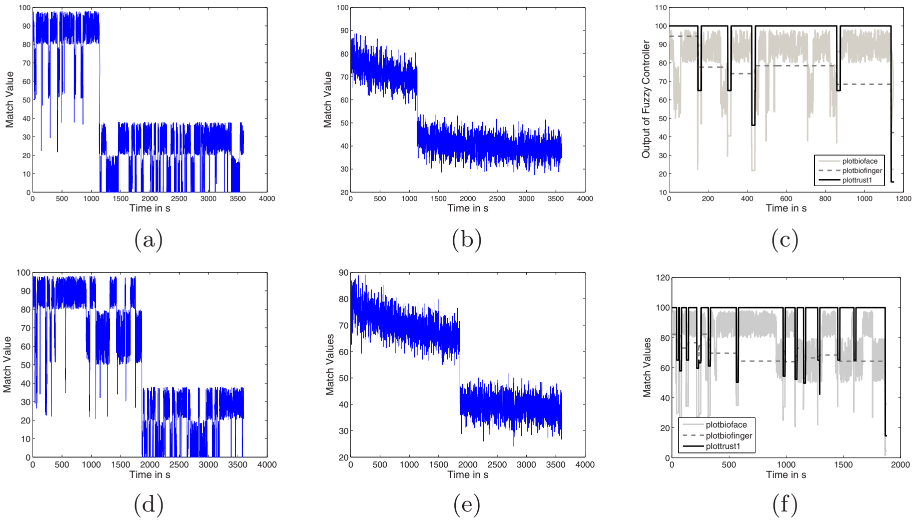
**Fig. 3.** Part of one genuine and impostor distribution of the simulated face biometric system

For all the experiments conducted the parameters values, related to the thresholds of face and trustworthiness, have been maintained to the best settings obtained in our previous work [3], in order to validate the generalization capabilities of the approach. The face threshold assumed values in the range $[0 - 100]$, while for TRUST a reduced range of $[50 - 70]$ has been considered, in order to ensure at least a satisfactory value of the trustworthiness during the entire working session. In particular the experiments demonstrate that is possible to effectively design a satisfactory balance between the number of finger requests and the trustworthiness by setting the parameters of the approach respectively to 50 for the face threshold $th$ and to 65 for the TRUST.

As previously indicated [3], another important aspect that can influence the number of the finger authentication requests during a working session regards the choice of the membership functions used by the fuzzy controller during the biometrics and trust evaluation. For this reason several experiments have been carried out in order to set, given the optimal values for the biometric parameters, the membership functions of the fuzzy controller that minimize the finger authentication requests during a working session, but paying more attention to the cases in which a relevant trustworthyness decrease unexpectedly occurs.

The simulations of this multi-modal authentication approach to one hundred of malicious users have been carried out, showing the behavior of the fuzzy controller for each user. An example is given in Figure 4, with the distributions of two users of the matching values of the malicious face (a),(d), and finger (b),(e), and the outputs obtained from the multi-modal fuzzy controller, that show the distributions of the TRUST, BIOFACE and BIOFINGER (c),(f).

It is important to underline that each user of the considered dataset is defined with different biometric matching scores, and for each user an independent impostor tries to attack the working session. Experiments show that the number of finger requests strongly depends on the behavior of the user in front of the camera, but when a malicious user attempts to corrupt the system, the trust value immediately undergoes the low level value and the impostor will be flush out by the multi-modal fuzzy controller, as shown in Figure 4(c) and (f).



(a) (b) (c)

(d) (e) (f)

**Fig. 4.** Distribution of face and finger of one impostor and output of the fuzzy controller

Furthermore, the entire simulation process shows how, for each user considered, the mean number of finger requests after a genuine-impostor swap is very low, equal to 0.84. In the best cases the finger request is not needed, since the low level trustworthiness threshold ends the suspect working session. The best results obtained from all the individuals during a malicious working session shows how our approach is able to ensure a continuous user authentication process and it is able to grant a satisfactory level of trustworthiness without interrupting the user too many times, capturing any intrusion tentative from an impostor user. Moreover, in [4] experimental results showing the good performance of our system during a complete working session without intrusion are presented.

## 4 Conclusions

In this paper we discuss a biometric approach for continuous user authentication by using a fuzzy controller. In particular we investigate the behavior of a multimodal biometric system as input designed with the aim of preventing user substitution after the initial authentication process. Satisfactory results obtained from the experiments show that the proposed approach for fuzzy continuous authentication is feasible and effective, flushing out the malicious attack.

## Acknowledgments

This work was partly funded by the European Community under the SecureSCM (Secure Supply Chain Management) contract AOR 213531,FP7-ICT-2007-1. The authors would also thank Roberto Sassi and Fabio Scotti for their support.

## References

1. Azzini, A., Damiani, E., Marrara, S.: Ensuring the identity of a user in time: a multi-modal fuzzy approach. In: CIMSA IEEE, Ostuni, Italy, June 27-29 (2007)
2. Azzini, A., Marrara, S.: A Fuzzy Trust Model Proposal to Ensure the Identity of a User in Time. In: Proceedings of the 9th Fuzzy Days, Dortmund, Germany, September 18-20 (2006)
3. Azzini, A., Marrara, S., Sassi, R., Scotti, F.: A Fuzzy Approach to Multimodal Biometric Authentication. In: Apolloni, B., Howlett, R.J., Jain, L. (eds.) KES 2007, Part II. LNCS (LNAI), vol. 4693, pp. 801–808. Springer, Heidelberg (2007)
4. Azzini, A., Marrara, S., Sassi, R., Scotti, F.: Multimodal User Authentication through Fuzzy Techniques. Technical Internal Report, 112 (May 2008)
5. Klir, G.J., Yuan, B.: Fuzzy Sets and Fuzzy Logic: Theory and Applications. Prentice Hall, Upper Saddle River (1995)
6. Bolle, R.M., Connell, J.H., Pankanti, S., Ratha, N.K., Senior, A.W.: Guide to Biometrics. Springer, New York (2004)
7. Cappelli, R., Maio, D., Maltoni, D., Wayman, J., Jain, A.: Performance Evalutation of Fingerprint Verification System. IEEE Transactions on pattern Analysis and Machine Intelligence 28(1) (2006)
8. Chang, K.I., Bowyer, K.W., Flynn, P.J.: An Evaluation of Multimodal 2D+3D Face Biometrics. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(4), 619–624 (2005)
9. Golfarelli, M., Maio, D., Maltoni, D.: On the Error-Reject Tradeoff in Biometric Verification Systems. IEEE Transactions on Pattern Analysis and Machine Intelligence
10. Maltoni, D., Maio, D., Jain, A.K., Prabhakar, S.: Handbook of Fingerprint Recognition. Springer, New York (2003)
11. Mamdani, E.H., Assilian, S.: An experiment in linguistic syntesis with a fuzzy logic controller. International Journal Man-Machine Studies 7, 1–13 (1975)
12. Schmidt, S., Steele, R., Dillon, T.: Towards Usage Policies for Fuzzy Inference Methodologies for Trust and QoS Assessment. In: Proceedings of Fuzzy Days 2006, Dortmund, Germany (September 2006)
13. Lee, S.-W., Li, S.Z.(eds.): Advances in Biometrics. Proceedings of ICB 2007, vol. 4642. Springer, Heidelberg (2007)
14. Snelick, R., Uludag, U., Mink, A., Indovina, M., Jain, A.K.: Large Scale Evaluation of Multimodal Biometric Authentication Using State-of-the-Art Systems. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(3), 450–455 (2005)
15. Uludag, U., Pankanti, S., Prabhakar, S., Jain, A.K.: Biometric Cryptosystems: issues and challenges. Proceedings of the IEEE 92(6), 948–960 (2004)

# A Learning Automata Approach to Multi-agent Policy Gradient Learning

Maarten Peeters[1,⋆], Ville Könönen[2], Katja Verbeeck[3], and Ann Nowé[1]

[1] Vrije Universiteit Brussel
Computational Modeling Lab
Pleinlaan 2
1050 Brussel, Belgium
[2] VTT Technical Research Centre of Finland
P.O. Box 1100
FI-90571 Oulu, Finland
[3] KaHo Sint-Lieven
Information Technology Group
Gebroeders Desmetstraat 1, B-9000 Gent, Belgium

**Abstract.** The policy gradient method is a popular technique for implementing reinforcement learning in an agent system. One of the reasons is that a policy gradient learner has a simple design and strong theoretical properties in single-agent domains. Previously, Williams showed that the REINFORCE algorithm is a special case of policy gradient learning. He also showed that a learning automaton could be seen as a special case of the REINFORCE algorithm. Learning automata theory guarantees that a group of automata will converge to a stable equilibrium in team games. In this paper we will show a theoretical connection between learning automata and policy gradient methods to transfer this theoretical result to multi-agent policy gradient learning. An appropriate exploration technique is crucial for the convergence of a multi-agent system. Since learning automata are guaranteed to converge, they posses such an exploration. We identify the identical mapping of a learning automaton onto the Boltzmann exploration strategy with an suitable temperature setting. The novel idea is that the temperature of the Boltzmann function is not dependent on time but on the action probabilities of the agents.

## 1 Introduction

In policy gradient learning, the parameters of the policy function get updated towards the gradient of the performance of the current policy. Their performance can even be augmented by means of approximating the value function, leading to theoretical convergence guarantees in single-agent MDPs [1,2]. While these policy gradient learners perform well in a wide variety of reinforcement learning tasks, there are no general theoretical guarantees how they will perform in a

⋆ Research funded by a Ph.D grant of the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT- Vlaanderen).

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 379–390, 2008.

multi-agent setting. However, previous research does indicate that policy gradient learners can perform well in multi-agent settings; in [2,3] the focus was on the convergence in a very simple 2-agent, 2-action setting, while in [4,5,6] settings where agents can observe the environment and the other agents were studied.

Research in the domain of learning automata theory began with the research of Tsetlin [7] in the beginning of the 60s. A Learning Automaton can be described as an independent decision making device that is suited for learning optimal control based on a scalar reinforcement signal. In its current form, a learning automaton models the internal state of an agent as a probability distribution according to which actions should be chosen [8]. These probabilities are adjusted according to some reinforcement scheme using the success or failure repsonse the environment generated. These schemes often have nice theoretical properties which lead to understanding their behavior in a multi-agent setting. An overview of the theory of learning automata is provided in [9] or more recently in [8].

An important topic in the field of LA is how a collection of automata could be interconnected and how these interconnections behave in a single environment. Even in a multi-agent environment they provide theoretical guarantees for convergence to an equilibrium point in a team game. Williams formulated the connection between a learning automaton using the $L_{R-I}$ algorithm and his REINFORCE algorithm [10]. Yet he only showed this connection for a 2-action learning automaton. In this paper we will construct a mapping between a policy gradient learner and an $n$-action learning automaton.

Exploration is a crucial factor for the convergence of multiple agents learning in the same environment. Since the learning automata are guaranteed to converge in a multi-agent setting, they posses such an essential exploration strategy. Because the Boltzmann function is a popular action selection strategy in multi-agent systems, we also investigate the mapping of the action selection strategy of learning automata onto the Boltzman function. The exploration of the learning automata depends purely on the action distribution over its actions. When mapping this exploration strategy we need to identify the appropriate temperature of the Boltzmann function for a given input.

The paper is structured as follows. We shortly repeat Markov Decision Processes and Policy gradient Learning in Sections 2 and 3 respectively. In Section 4 we give a short overview of learning automata theory. Next, in Section 5 we proof there is a theoretical connection between learning automata and the policy gradient method. In the section thereafter we identify the mapping between the exploration of the learning automata and the temperature for the identical Boltzmann function. We show by empirical results that the exploration strategy we found outperforms a Boltzmann with a decaying temperature (which is often used in multi-agent learning research). In the final section we summarize and conclude.

## 2 Situated Learner

In this paper we consider the standard Markov Decision Process (MDP) and reinforcement learning framework as can be found in the literature [11,12]

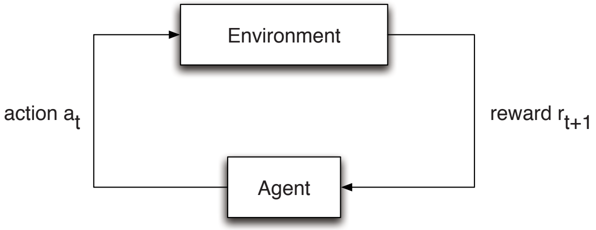


**Fig. 1.** Learning System in its environment

(schematically depicted in Figure 1). We consider an agent to be situated in an environment. We assume that time progresses in discrete time steps: $t \in \{1, 2, \dots\}$. The state the agent is situated in is $s_t \in S$ and the action it takes is $a_t \in A$. The dynamics of the environment are defined by the state transition probabilities: $P^a_{ss'} = Pr\{s_{t+1} = s' | s_t = s, a_t = a\}$ and the expected rewards: $R^a_s = E\{r_{t+1} | s_t = s, a_t = a\}$, $\forall s, s' \in S$ and $\forall a \in A$. How an agent selects its next action is defined by its policy function $\pi(s, a, \boldsymbol{\theta})$. The only restriction we put on this policy function is that it is differentiable with respect to its parameters, i.e. $\frac{\partial \pi(s,a,\boldsymbol{\theta})}{\partial \theta}$ exists for all $\theta$.

In this paper the set of actions is assumed to be discrete. Further, we use a subscript for indicating time steps, e.g. $a_t$ is the action selected by the agent at the time step $t$. The reward signal, resulting from action $a_t$ being selected at time instance $t$ is denoted by $r_{t+1}$. The goal of the agent can be described as optimizing the long-term expected discounted reward from starting state $s_0$:

$$\rho(\pi, s_0) = \lim_{t \to \infty} E\{r_1 + \gamma r_2 + \cdots + \gamma^{t-1} r_t | \pi, s_0\} = \sum_a \pi(s_0, a, \boldsymbol{\theta}) Q^\pi(s_0, a), \quad (1)$$

where the state-action pair values for a certain policy are given by:

$$Q^\pi(s, a) = \lim_{t \to \infty} E\{r_1 + \gamma r_2 + \cdots + \gamma^{t-1} r_t | \pi, s_0 = s, a_0 = a\} \quad (2)$$

## 3 Policy Gradient Learning

Function approximation has always been a fundamental element of reinforcement learning theory. However, most of the research focused on finding a good approximation of the *value function* and then compute the policy using the value function approximation. While this approach provides a good solution to many problems, it has its limitations. It is oriented towards finding deterministic policies and arbitrarily small changes in the estimated value of an action can give discontinous changes in the action selection mechanism. The latter turned out to be a key problem for establishing convergence guarantees for value-function based algorithms.[1].

Another approach is to approximate the policy function directly using an independent function approximator with its own parameters. This approximator could be a neural network that outputs action probabilities, with the parameters to be updated being the weights of the nerural network. Formally we can say that if $\boldsymbol{\theta}$ are the parameters of the policy $\pi$ and $\rho$ is the expected discounted reward for policy $\pi$, then in the policy gradient approach, the parameters of $\pi$ should be updated according to:

$$\Delta\theta \propto \gamma \frac{\partial \rho}{\partial \theta}, \tag{3}$$

where $\gamma$ is the learning rate or step size. If this condition is met, then $\boldsymbol{\theta}$ can be assured to converge to a locally optimal solution in a single agent MDP. In [2] it was proven that an unbiased estimate of the gradient can be obtained from experience using an approximate value function satisfying certain properties.

## 4 Learning Automata

A learning automaton is an autonomous mechanism that is capable of making adaptive decisions in a highly uncertain, stochastic environment. Its goal is to select the optimal action from the set of possible actions based on reinforcement signals given by the environment. We start the section by discussing the formal model of the agent–environment interaction in general and the learning automata learning model in particular and then we proceed to the learning algorithms used with the model.

### 4.1 Formal Model for Environment–Automaton Interaction

An automaton, situated in an environment, selects at each time step, one action $a$ from the set of all possible actions $A$ using an internal action probability vector $\mathbf{p}$, i.e. $p(a) \in [0,1], \forall a \in A$ and $\sum_a p(a) = 1$. The probability that the learning automaton selects action $i$ is thus given by $p(i)$. In fact, $\mathbf{p}$ directly defines the policy. Based on this action selection, the environment then gives a reward or feedback signal $r$ to the agent. The feedback signal is assumed to be binary $r \in \{0,1\}$, where 1 denotes a positive feedback from the environments and 0 is a negative feedback. In addition, the signal is stochastic, i.e. the environment has a distribution $\mathbf{c}$ with $c(a) = Pr\{r = 1|a\}$. Note that in the description of a learning automaton we did not talk about the concept of states. This is because in its purest, simplest form a learning automaton is stateless and as such has no notion of states.

Learning automata utilize the above described model for agent–environment interaction for learning optimal performance of the agent in unknown, stochastic environments. Based on the feedback signal, the probability distribution $\mathbf{p}$ is updated each time step by using an updating scheme. Different update schemes with varying properties exist. In this paper, we apply the *Linear Reward–Inaction* ($L_{R-I}$) scheme that is described next.

### 4.2 Linear Reward–Inaction Scheme

The idea behind this update scheme is that when an action was successful, i.e. $r = 1$, the action probability for this action should be increased and all other action probabilities should be decreased appropriately. However, when an action was not successful, the action probabilities remain the same.

$$\begin{aligned} a_t &= a \text{ i.e. action selected at timestep } t \\ p_{t+1}(a) &= p_t(a) + \alpha_t r_{t+1}[1 - p_t(a)] \\ p_{t+1}(b) &= p_t(b) - \alpha_t r_{t+1} p_t(b), \forall b \neq a, \end{aligned}$$

where $\alpha_t \in [0, 1]$ is the step size or learning rate parameter. We can rewrite these equations using vector notation as follows:

$$\mathbf{p}_{t+1} = \mathbf{p}_t + \alpha_t r_{t+1}[\mathbf{e_a} - \mathbf{p_t}], \tag{4}$$

where $\mathbf{e_a}$ is defined as a unit vector with unity at position $a$ which corresponds to the action select at time $t$. In our experiments $r_{t+1}$ is a binary valued variable while the learning automata theory is provided for a reward signal from the continuous interval $[0, 1]$. This assumption can be made without loss of generality.

An important result for the linear-reward inaction scheme is the following:

**Theorem 1.** [9] *A team of learning automata using the $L_{R-I}$ update scheme with a suitable step size will independently of each other (i.e. without any communication) converge to a pure Nash equilibrium point in a common pay-off game (i.e. team game)*

## 5 A Learning Automaton as a Policy Gradient Learner

Williams showed how a 2-action learning automaton using the $L_{R-I}$ scheme is a special case of the REINFORCE algorithm [10]. In this section we will give the theoretical connection between the policy gradient method and the more general $n$-action learning automaton. One of the main differences between REINFORCE and the general policy gradient method is that REINFORCE only uses the actual returns to find an unbiased estimate of the gradient, while the general policy gradient method use the assistance of a learned value-function. In Section 6 this will be one the of the restrictive conditions we will try to relax.

In order to formalise the mapping, we need to find the link between the different components of both mechanisms such as the policy, the update method and the learning rate. First note that the parameter vector $\boldsymbol{\theta}$ for a learning automaton is exactly the action probability vector $\mathbf{p}$, which also represents the current stochastic policy. So this means that the parametrization of the policy here is given by the identical function:

$$\pi(a, \boldsymbol{\theta}) = \boldsymbol{\theta}^T \mathbf{e_a} = \mathbf{p}^T \mathbf{e_a}.$$

If we let $\delta_i(a)$ be the Kronecker delta defined as

$$\delta_i(a) = \begin{cases} 1 & \text{if a is the } i^{th} \text{ action} \\ 0 & \text{otherwise,} \end{cases}$$

Then we have that the derivative of the policy towards one of its parameters for a learning automaton is exactly the Kronecker delta function:

$$\frac{\partial \pi(a, \boldsymbol{\theta})}{\partial \theta^i} = \delta_i(a)$$

According to the policy gradient method the parameters of the policy function are updated as follows:

$$\theta^i_{t+1} = \theta^i + \gamma^i \sum_a \frac{\partial \pi(a, \boldsymbol{\theta}, s_0)}{\partial \theta^i} Q^\pi(s_0, a) \tag{5}$$

In case of the stateless learning automaton, this update becomes:

$$\begin{aligned} \theta^i_{t+1} &= \theta^i + \gamma^i \sum_a \frac{\partial \pi(a, \boldsymbol{\theta})}{\partial \theta^i} Q^\pi(a) \\ &= \theta^i_t + \gamma^i \sum_a \delta_i(a) Q^\pi(a) \end{aligned} \tag{6}$$

where $\gamma^i$ is the learning rate for the $i^{th}$ action. However, a learning automaton does not store any approximation of action values or average rewards obtained for an action. The stateless automaton only uses the actual returns to estimate $Q^\pi(a)$.This results in:

$$\begin{aligned} \theta^i_{t+1} &\approx \theta^i_t + \gamma^i \sum_a \delta_i(a) r_{t+1} \\ &= \theta^i_t + \gamma^i r_{t+1} \end{aligned}$$

If we write this in vector notation we get:

$$\boldsymbol{\theta}_{t+1} \approx \boldsymbol{\theta}_t + \boldsymbol{\gamma} r_{t+1} \tag{7}$$

with $\boldsymbol{\gamma}$ the vector of learning rates, one for each $\theta^i$.

Now, the only step that remains is to identify the connection between the updating of the automaton according to the learning automaton formulation given in Equation 4 and the policy gradient view given in Equation 7. This leads to :

$$\begin{aligned} \boldsymbol{\gamma}\ r_{t+1} &= \alpha_t r_{t+1} [\mathbf{e_a} - \mathbf{p}_t] \\ \boldsymbol{\gamma} &= \alpha_t [\mathbf{e_a} - \mathbf{p_t}] \end{aligned} \tag{8}$$

Thus we can conclude that a learning automaton is a stateless policy gradient learner which uses the identical function as parametrization of its policy parameters and uses a variable step size equal to the one given in Equation 8.

As a consequence, many results of learning automata theory, such as Theorem 1, carry over directly to policy gradient methods (under the constraints explained above). Thus independent policy gradient learners can now be used to play stateless games under the guarantee of convergence to a Nash equilibrium.

# 6 Experiments: Multi-agent Policy Gradient Learning

Learning in a multi-agent environment is substantially more difficult than learning in a single agent environment. Because multiple agents are acting in the same environment, this environment is perceived as non-stationary from the viewpoint of each agent. Due to this non-stationarity, convergence guarantees of traditional single-agent learning techniques become invalid or the single-agent algorithms malfunction. Furthermore, characteristics like asynchronous action selection, delayed rewards, incomplete information and conflicting interests come into play. All these characteristics make learning in a multi-agent system a very challenging task.

In this section, we will test the performance of the policy gradient method and the learning automata in a single-state game. In the game, there are two players, player 1 (the row player) and agent 2 (the column player). Both agents share the same utility function, i.e. the game is a team game. In the first set of experiments, the settings of the policy gradient method will be those that were established in Section 5. In the next experiments we relaxed these conditions to investigate their influence on the performance. We use two performance criteria: 1) the number of times the method finds the optimal solution 2) the speed of convergence.

## 6.1 Test Games

The test game we used is shown in Fig. 1 on the left and is called the Climbing Game (originally published in [13]).

In the Climbing Game, the optimal solution is attained when both agents select their $1^{st}$ action. However, large negative rewards surround the optimal solution and therefore it might be more favorable to select another safer, yet sub-optimal action.

Since we are going to apply the learning automata update scheme of Equation 4 , we first need to normalize the rewards in the unit interval and then use those values stochastically, i.e. the reward is used as a probability of receiving a positive reward signal equal to 1. So this means that rewards are actually generated with a binomial distribution whose mean is dictated by the game.

## 6.2 Applying Learning Automata Theory

In the previous section we have shown how a learning automaton can be seen as policy gradient learner with a variable learning rate. In the experiment we let

**Table 1.** The single-state games used in the test runs. The game on the left is the Climbing Game and the game on the right is a 2 player variant. In both cases, both agents share the same utility function.

$$\left(\begin{array}{c|ccc} & b_1 & b_2 & b_3 \\ \hline a_1 & 11,11 & -30,-30 & 0,0 \\ a_2 & -30,-30 & 7,7 & 6,6 \\ a_3 & 0,0 & 0,0 & 5,5 \end{array}\right) \left(\begin{array}{c|cc} & b_1 & b_2 \\ \hline a_1 & 100,100 & 99,99 \\ a_2 & 0,0 & 98,98 \end{array}\right)$$

**Table 2.** Percentage of convergence and average reward for a pair of learning automata and a pair of policy gradient learners using the identical mapping playing the Climbing game (averaged over 1000 runs)

| Climbing Game | | | | | |
|---|---|---|---|---|---|
| Algorithm | $\alpha$ | % NE | % Opt | Avg. Rew | Avg. Steps |
| LA | 0.01 | 17.4 | 9.6 | 6.236 | 8972.28 |
| LA | 0.05 | 32.4 | 21.2 | 6.75 | 812.834 |
| LA | 0.1 | 39.2 | 25.4 | 6.632 | 275.796 |
| LA | 0.25 | 33.4 | 22 | 5.42 | 61.988 |
| LA | 0.5 | 32.8 | 16 | 4.8 | 20.342 |
| PG as LA | 0.01 | 18.8 | 9.8 | 6.262 | 9148.08 |
| PG as LA | 0.05 | 34.4 | 21.2 | 6.702 | 818.38 |
| PG as LA | 0.1 | 37.4 | 24.8 | 6.336 | 262.39 |
| PG as LA | 0.25 | 38.4 | 22.4 | 5.588 | 61.478 |
| PG as LA | 0.5 | 33.4 | 18 | 4.864 | 20.124 |

two automata (or two agents) play the climbing game for at least 5000 iterations. After these 5000 training iterations we keep a log of which action each of the two players converged to. The results for the climbing game are shown in Table 2. We compute the percentage of convergence to the optimal action and the average reward obtained all averaged over 1000 games. Although learning automata theory guarantees convergence to a Nash equilibrium, the empirical results do not seem to comply with this theory. This can be explained that the theoretical guarantee is only valid under an adequate choice of learning rate parameter. An experiment with a learning rate setting of $\alpha = 0.00005$ resulted in a convergence of over 90% to a Nash equilibrium.

The following conclusions can be drawn from the table. Both techniques have almost the same percentage of finding a Nash equilibrium and the optimal joint-action. Also, the average reward tells us that when the agents do not converge to a Nash equilibrium, they converge to the same sub-optimal actions. Thus the performance of both techniques are almost identical. Which is not surprising given the above mapping. This conclusion can be strengthened by two statistical tests.

For a first test, we investigate the relative frequencies of the possible joint-actions. The central limit theorem says that each relative frequency $p_{ij}$ (i.e. the absolute frequencies, $q_{ij}$, divided by the number of iterations) is distributed according to a normal distribution: $p_{ij} \sim N(q_{ij}, \frac{q_{ij}(1-q_{ij})}{n})$ and the difference of two such distributions is again a normal distribution: $p^I_{ij} - p^{II}_{ij} \sim N(q^I_{ij} - q^{II}_{ij}, \frac{q^I_{ij}(1-q^I_{ij})}{n} + \frac{q^{II}_{ij}(1-q^{II}_{ij})}{n})$. We can now compute the statistic $T$ as

$$T = \sum_i^{rows} \sum_j^{cols} \frac{(q^I_{ij} - q^{II}_{ij})^2}{\frac{q^I_{ij}(1-q^I_{ij})}{n} + \frac{q^{II}_{ij}(1-q^{II}_{ij})}{n}}. \tag{9}$$

Since we assume both algorithms to be identical ($H_0$ can be formulated as: *we detect no statistical difference between both algorithms*) we assume that Equation

9 is distributed according to the $\chi^2$ distribution with one degree of freedom less than the number of categories. If this assumption is wrong, $T$ will be larger than the critical value found by the $\chi^2$ percentile point and we can reject $H_0$ (which would indicate that the algorithms differ). We computed different $T$ values for different learning rates, see the left column of Table 3. The critical value we used is $\chi^2_8$ because we have 9 different joint-actions but one joint-action can be computed as one minus the sum of all the others which leaves us with one degree of freedom less. The statistic $T$ is smaller then the critical value for all the tested settings of learning automata. We can conclude that there is no statistical difference between the two algorithms.

A second statistical test we applied on our data is Bowker's Test for Symmetry [14]. This test not only compares the performance of the algorithms but it also considers the time steps at which the algorithms converged to the same solution. Bowker's test is thus a stronger and more restrictive test.

**Table 3.** Statistics obtained for the chi-square test and Bowker's Test for Symmetry playing the Climbing Game. All statistics are tested against a $\chi^2$ distribution with an accuracy of 95%.

| $\alpha$ | $T$ | $\alpha$ | $Q_{Bowker}$ |
|---|---|---|---|
| 0.01 | 3.919728 | 0.01 | 10.82503 |
| 0.05 | 5.636051 | 0.05 | 14.82659 |
| 0.1 | 7.605683 | 0.1 | 19.29709 |
| 0.5 | 6.936435 | 0.5 | 28.06774 |
| | compare to $\chi^2_8 = 15.50731$ | | compare to $\chi^2_{36} = 50.99846$ |

All the results in the table confirm what we expected: there is no statistical difference detected between the learning automata and the policy gradient learners with the same parameterization and the same learning rate even in the more restrictive Bowker's test.

### 6.3 A Different Policy Parameterization

In the next experiment the policy parametrization function that is used is no longer the identical function. Instead we use a Boltzmann distribution as a parametrization function:

$$\pi(a, \boldsymbol{\theta}) = \frac{e^{\boldsymbol{\theta}' \mathbf{e_a}/T}}{\sum_{b \in A} e^{\boldsymbol{\theta}' \mathbf{e_b}/T}} \tag{10}$$

with a temperature setting value of T denoting the amount of exploration.

Note that the Boltzmann distribution is a smooth function with respect to its parameters $\boldsymbol{\theta}$. Because of this choice for the policy function, the policy parameters $\boldsymbol{\theta}$ do not represent action probabilities anymore. Therefore the theoretical connection is no longer fulfilled. Furthermore, we will use a value-function approximation $f_{\boldsymbol{\omega}}$ that stores an indication of the quality of each action. The parameters $\boldsymbol{\omega}$ are updated according to:

$$\omega_{t+1}(a) = (1-\beta)\omega_t(a) + \beta r_{t+1}, \tag{11}$$

where $\beta$ is the learning rate parameter. The above learning rule is guaranteed to converge to the optimal value function if all actions are tried infinitely and $\beta$ fulfills the following conditions[2]. The learning rate $\beta$ should be taken in $[0,1]$ and $\sum_t^\infty \beta_t = \infty$ and $\sum_t^\infty \beta_t^2 < \infty$. This approximation is then used for updating the parameters of the policy function as is expressed in Equation 6.

As pointed out before, the amount of exploration is a crucial factor in the convergence of a multi-agent system. Learning automata apply the action probabilities directly without the use of an action selection function (i.e. they use the identical function). If we want to map this behavior for 2 actions onto a Boltzmann exploration function, we are looking for the temperatures $t_1$ and $t_2$ that give fix-points of the equations

$$\begin{cases} p_1 = & \dfrac{e^{\frac{p_1}{t_1}}}{e^{\frac{p_1}{t_1}} + e^{\frac{p_2}{t_2}}} \\ p_2 = & \dfrac{e^{\frac{p_2}{t_2}}}{e^{\frac{p_1}{t_1}} + e^{\frac{p_2}{t_2}}} \end{cases} \tag{12}$$

For 2 actions the temperatures $t_1$ and $t_2$ are identical and can be expressed as: $t = \frac{p_1 - p_2}{\ln p_1 - \ln p_2}$. We tested this new setting on a simple 2 player, 2 action game (Figure 1).

**Table 4.** Experimental results of the Policy gradient method using Boltzmann exploration playing the 2 action climbing game. The table shows the results for different temperature setting. The first part shows the results for the temperature setting based on Equation 12. Part 2–4 shows the results for decaying temperature settings with different starting temperatures.

| PGM - Boltzmann<br>action prob. based temperature | | | |
|---|---|---|---|
| $\alpha$ | % NE | AvgRew | AvgSteps |
| 0.01 | 22 | 99.11 | 4884.91 |
| 0.05 | 37 | 99.36 | 3520.98 |
| 0.25 | 28 | 99.24 | 1310.21 |
| 0.1 | 36 | 99.33 | 589.64 |
| 0.5 | 31 | 99.14 | 196.93 |

| PGM - Boltzmann<br>Decaying temperature | | | | | |
|---|---|---|---|---|---|
| startT | decay | $\alpha$ | % NE | AvgRew | AvgSteps |
| 2 | 0.9994 | 0.01 | 0 | 99 | 3847.91 |
| 2 | 0.9990 | 0.05 | 0 | 99 | 1942.92 |
| 2 | 0.9984 | 0.1 | 0 | 99 | 1219.63 |
| 2 | 0.9963 | 0.25 | 3 | 99.03 | 543.21 |
| 2 | 0.9951 | 0.5 | 12 | 99.07 | 148.7 |

| PGM - Boltzmann<br>Decaying temperature | | | | | |
|---|---|---|---|---|---|
| startT | decay | $\alpha$ | % NE | AvgRew | AvgSteps |
| 10 | 0.9991 | 0.01 | 0 | 99 | 3614.71 |
| 10 | 0.9985 | 0.05 | 0 | 99 | 1525.63 |
| 10 | 0.9975 | 0.1 | 0 | 99 | 1055.09 |
| 10 | 0.9943 | 0.25 | 0 | 99 | 464.27 |
| 10 | 0.9772 | 0.5 | 0 | 99 | 130.83 |

| PGM - Boltzmann<br>Decaying temperature | | | | | |
|---|---|---|---|---|---|
| startT | decay | $\alpha$ | % NE | AvgRew | AvgSteps |
| 0.5 | 0.9997 | 0.01 | 0 | 99 | 4919.91 |
| 0.5 | 0.9995 | 0.05 | 2 | 99.02 | 2677.02 |
| 0.5 | 0.9991 | 0.1 | 11 | 99.1 | 1558.28 |
| 0.5 | 0.9963 | 0.25 | 5 | 99.05 | 542.05 |
| 0.5 | 0.9851 | 0.5 | 10 | 99.1 | 147.12 |

We tested 4 different policy gradient learners on this game. The first one is the policy gradient learner using Boltzmann as its action selection strategy with a temperature set by Equation 12. The second to fourth results show policy gradient learners also using the Boltzmann function as their action selection, however, they use a temperature starting at *startT* which is decayed with a factor *decay* every timestep. The results show that the policy gradient learner using the temperature that doesn't depend on time outperforms the other learners.

## 7 Conclusion

In this paper we investigated the connection between a policy gradient learer and learning automaton. We showed theoretically and empirically that a policy gradient learner with an identical function parametrization and a variable learning rate matches the behaviour and characteristics of a learning automaton. By establishing this connection we argued that the theoretical properties of a multi-automata system; among which is the guarantee to convergence to a Nash equilibrium without any communication in a team game, should be transferable to independent policy gradient learners.

Furthermore we found an expression for the temperature for finding the fixpoint of the Boltzmann exploration function. This temperature is no longer dependent on the time but on the action probabilities of the agent. We showed empirically that policy gradient learners using this novel temperature setting have higher percentage of convergence to a Nash equilibrium.

In future research we will further try to adopt the theoretical properties of learning automata to suit the framework of policy gradient learners. This should give us a broader base for relaxing the necessary conditions for theoretical convergence and provide a major contribution to the field of policy gradient learning. A second track that remains to be investigated is the application of the variable learning rate to more general reinforcement learners combined with the new exploration strategy.

## References

1. Bertsekas, D.P., Tsitsiklis, J.N.: Neuro-Dynamic Programming. Athena Scientific (1996)
2. Sutton, R.S., McAllester, D., Singh, S.P., Mansour, Y.: Policy gradient methods for reinforcement learning with function approximation. In: Advances in Neural Information Processing Systems (NIPS 1999), Denver, CO, pp. 1057–1063 (2000)
3. Bowling, M., Veloso, M.M.: Multiagent learning using a variable learning rate. Artificial Intelligence 136(2), 454–460 (2002)
4. Peshkin, L., Kim, K.E., Meuleau, N., Kaelbling, L.P.: Learning to cooperate via policy-search. In: Proceedings of the Sixteenth Conference on Uncertainty in Artifical Intelligence (UAI 2000), Stanford, CA, pp. 489–496 (2000)
5. Könönen, V.: Gradient based method for symmetric and asymmetric multiagent reinforcement learning. Web Intelligence and Agent Systems: An International Journal (WIAS) 3(1), 17–30 (2005)

6. Könönen, V.: Multiagent Reinforcement Learning in Markov Games: Asymmetric and Symmetric Approaches. PhD thesis, Helsinki University of Technology, Helsinki, Finland (2004)
7. Tsetlin, M.L.: Automata Theory and Modeling of Biological Systems. Academic Press, New York (1973)
8. Thathachar, M.A.L., Sastry, P.S.: Networks of Learning Automata: Techniques for Online Stochastic Optimization. Kluwer Academic Publishers, Dordrecht (2004)
9. Narendra, K.S., Thathachar, M.A.L.: Learning Automata: An Introduction. Prentice Hall, Englewood Cliffs (1989)
10. Williams, R.J.: Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine Learning 8(3–4) (1992)
11. Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction. MIT Press, Cambridge (1998)
12. Filar, J.A., Vrieze, K.: Competitive Markov Decision Processes. Springer, Heidelberg (1997)
13. Claus, C., Boutilier, C.: The dynamics of reinforcement learning in cooperative multiagent systems. In: Proceedings of the Fifteenth National Conference of Artificial Intelligence (AAAI 1998), Madison, WI, pp. 746–752 (1998)
14. Bowker, A.H.: Bowker's test for symmetry. Journal of the American Statistical Association 43, 75–83 (1984)

# Coordinated Exploration in Conflicting Multi-stage Games

Maarten Peeters[1,*], Ville Könönen[2], Katja Verbeeck[3], Sven Van Segbroeck[1], and Ann Nowé[1]

[1] Vrije Universiteit Brussel, CoMo Lab, Pleinlaan 2, 1050 Brussel, Belgium
[2] VTT Technical Research Centre of Finland, P.O. Box 1100, FI-90571 Oulu, Finland
[3] KaHo Sint-Lieven, Information Technology Group, Gebr. Desmetstraat 1, 9000 Gent, Belgium

**Abstract.** The analysis of the collective behavior of agents in a distributed multi-agent environment received a lot of attention in the past decade. More accurately, coordination was studied intensely because it enables agents to converge to Pareto optimal solutions and Nash equilibria. Most of these studies focussed on team games. In this paper we report on a technique for finding fair solutions in conflicting interest multi-stage games. Our hierarchical periodic policies algorithm is based on the characteristics of a homo egualis society in which the players also care about the proportional distribution of the pay-off in relation to the pay-off of the other players. This feature is built into a hierarchy of learning automata which is suited for playing sequential decision problems.

## 1 Introduction

A great deal of attention in the multi-agent learning community has been focussed on coordination because coordination enables agents to converge to desired equilibrium points; often Nash equilibria which aren't always fair solutions. These equilibrium points might be points where all the agents receive their highest possible pay-off (i.e. fair) or where only one agent receives a high pay-off (i.e. possible unfair). Another performance criteria is how well the overall system behaves. For some systems, a fair criteria would be to take the performance of the worst performing agent as the performance of the global system. Thus the faster *all* machines process their jobs, the better we grade the system. Defining what an optimal solution is for all agents is relatively easy in an identical pay-off game, however in this example the machines are also in competition for resources and as such, the game can be more naturally modeled as a conflicting interest game. Defining what the fairest solution is in a conflicting interest game is harder and designing agents that can learn to play such a policy is even more complex.

In this paper, we propose a technique called *hierarchical periodic policies* for a hierarchy of learning automata. The algorithm is based on excluding actions.

[*] Research funded by a Ph.D grant of the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT- Vlaanderen).

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 391–402, 2008.

By doing so, the agents shrink the joint-action space and the search for the optimum can be continued in this reduced joint-action space. However, since the agents are playing conflicting interest games, they have different preferences to different solutions in the game. Therefore, we have to incorporate a mechanism by which the agents will allow each other to visit different equilibria even if those equilibria are not personally favorable. This mechanism of social behavior is based on the characteristics of a homo egualis society [1]. In a homo egualis society, the players do not only care about their personal pay-off but also on how this pay-off relates to the pay-off of the other players.

The remainder of the paper is organised as follows. In the following section we show that pure and mixed Nash equilibria might not always be fair results. We also provide explain what how a fair solution can be found by alternating the policy that is played. Thereafter, we show how sequential decision problems can be modeled as multi-stage games. In Section 3 learning automata are introduced. Different learning automata can be combined to form more complex structures such as hierarchies. The following section explains the algorithm of the periodic policies. In Section 5 we report on related work which provided baseline results for our experiments. In a final section we conclude.

## 2 Conflicting Interest Games

### 2.1 Fairness

Consider the Battle of Sexes game (see Table 1). By analysing the game it is easy to see that the pure Nash equilibria are joint-actions *(Football, Football)* and *(Opera,Opera)*. Although both Nash equilibria are the only optimal pure solutions in the game, in terms of fairness, these solutions are not optimal since both Nash equilibria favor either one of the players.

Since the pure Nash equilibria are not fair, we could check the mixed Nash equilibria. It is easy to verify that the mixed Nash equilibrium is $((\frac{2}{3},\frac{1}{3}),(\frac{1}{3},\frac{2}{3}))$ Thus the row player selects action *Football* with probability $\frac{2}{3}$ and the action *Opera* with probability $\frac{1}{3}$. The column player selects action *Football* with probability $\frac{1}{3}$ and the action *Opera* with probability $\frac{2}{3}$. By playing this mixed strategy, both players receive an expected pay-off of $\frac{2}{3}$. This is indeed a fair solution, however, this pay-off is less than what both agents would receive by playing either of the pure Nash equilibria.

The solution concept called *periodic policies* [2] provides a technique for single state games by which the agents alternate between converging to different pure Nash equilibria in a manner such that the overall average pay-off of all the agents is both fair and as high as possible. In the example of the Battle of the Sexes

**Table 1.** The reward matrix of the Battle of the Sexes

$$M = \left( \begin{array}{c|cc} & \text{Football} & \text{Opera} \\ \hline \text{Football} & (2,1) & (0,0) \\ \text{Opera} & (0,0) & (1,2) \end{array} \right)$$

game, the players would alternate between the two pure Nash equilibria, giving an average reward of 1.5 for both players which is fair and significantly better compared to the pay-off of the mixed Nash equilibrium.

A periodic policy is defined as follows:

**Definition 1.** [3] *In a periodic policy, the agents alternate between periods of time in which different pure Nash equilibria are played.*

The inspiration for the periodic policies algorithm comes from the homo egualis society. In such a community, the players do not only care about their own personal gain or pay-off, but also how their own utility relates to the utilities of the other players.

### 2.2 Multi-stage Games

A multi-stage game is a game where the participating agents have to take a sequence of actions. A traditional MDP can be extended to the multi-agent case, called a Markov Game. Formally, we can define a Markov Game as a five-tuple, $M = \langle \mathbb{A}, \{A_i\}_{\forall i \in \mathbb{A}}, \mathbb{S}, T, R \rangle$ where:

- $\mathbb{A}$ is the set of agents participating in the game,
- $\{A_i\}_{i \in \mathbb{A}}$ is the sets of actions available to agent $i$,
- $S$ is the set of states (same as defined with an MDP),
- $T(s, \boldsymbol{a}, s')$ is the transition function stating the probability that a joint-action $\boldsymbol{a}$ will lead the agents from state $s$ to state $s'$,
- and $R^i : S \rightarrow \mathbb{R}$ is the reward function denoting the reward agent $i$ gets for entering a certain state.

In the remainder of this paper we limited ourselves to tree-structured multi-stage games. This means that there are no loops possible between the game stages and once branches are separated their paths will never be joined again.

We can view a multi-stage game as a sequence of single state games.

## 3 Learning Automata

### 3.1 Learning Automata Model

A learning automaton is an independent entity that is situated in a random environment and is capable of taking actions autonomously. The stochastic environment is responsible for generating a scalar value indicating the quality of the action taken by the learning automaton. This scalar value, which we call reward, is then fed back into the learning automaton.

The quadruple $\langle A, \mathbf{R}, \mathbf{p}, U \rangle$ expresses a learning automaton. $A = \{a(1), \ldots, a(n)\}$ denotes the set of actions the learning automaton can take. $R_t(a_t(i)) \in \mathbf{R}$ is the input that is given to the LA to indicate the quality of the chosen action. The probabilities of the automaton for selecting action $a_t(i)$ are stored in the vector $\mathbf{p_t} = [p_t(1), \ldots, p_t(n)]$. Note that at the beginning of the game all action probabilities are chosen to be equal: $p_0(1) = p_0(2) = \cdots = p_0(n) = \frac{1}{n}$. Each iteration the action probabilities are updated based on the reinforcement obtained from the environment using an update algorithm $U$. For the experiments in this paper, we used the Linear Reward-Inaction ($L_{R-I}$) [4] scheme because this scheme guarantees convergence to a pure policy in single-state games. Let $a_t = a(i)$ be the action chosen at time step $t$. Then the action probability vector $\mathbf{p}$ is updated according to

$$\mathbf{p_{t+1}} = \mathbf{p_t} + \alpha r_t(\mathbf{e_{a_t}} - \mathbf{p_t}) \tag{1}$$

with $\alpha$ the step size parameter and $\mathbf{e_{a_t}}$ a unit vector with unity at position $a_t$. Multiple automata can be combined into more complex structures such as hierarchies which are described next.

### 3.2 Hierarchical Learning Automata

The interaction of the two hierarchical agents in Figure 1 goes as follows. At the top level (or in the first stage) Agent 1 and Agent 2 meet each other in a game with stochastic rewards. They both take an action using their top level learning automata $LA\ A$ and $LA\ B$. Performing actions $a_i$ by $LA\ A$ and $b_k$ by $LA\ B$ is equivalent to choosing automata $LA\ A_i$ and $LA\ B_k$ to take actions at the next level. The response of environment $E_1$: $r_t \in \{0, 1\}$, is a success or failure, where the probability of success is given by $c^1_{ik}$ . At the second level the learning automata $LA\ A_i$ and $LA\ B_k$ choose their actions $a_{ij}$ and $b_{kl}$ respectively and these will elicit a response from environment of which the probability of getting a positive reward is given by $c^2_{ij,kl}$. At the end of the episode all the automata that were involved in one of the games, update their action selection probabilities based on the actions performed and the responses of the environments using the $L_{R-I}$ update scheme.

## 4 Learning Fair Policies

To obtain a fair solution as defined above in multi-stage conflicting interest games, we will include a social feature inspired by the homo egualis equation. By extending hierarchical learning automata with the ability to (temporarily) exclude actions, the joint-action space is reduced in size and the agents can find other equilibria. If the actions are excluded in an intelligent manner, the automata are able to find different periodic policies which favored either one of the players, resulting in an overall fair average.

This core idea of the technique is not new, in [5,2] an algorithm is introduced to let agents learn fair periodic policies in one-stage conflicting general-sum games.

Before we explain the algorithm we have to initialise the agents. In the initialization, we initialize the averages of the last phase ($w^h$) and the global average ($v^h$) to 0.

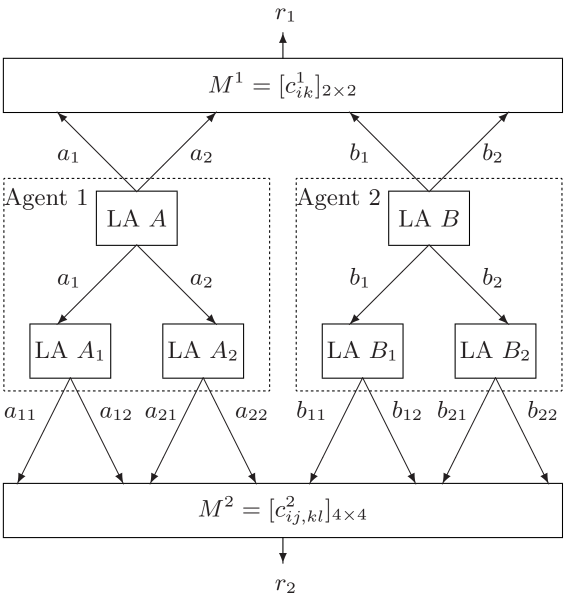


**Fig. 1.** An interaction of two agents constructed by learning automata hierarchies. The top-level automata play a single stage game and produce a reward $r_1$. Then one learning automata of each hierarchy at the second level play another single stage game, resulting in reward $r_2$.

The algorithm itself consists of two different phases, the independent exploration phase and the communication phase. During the independent learning phase the agents start to explore their action space without any form of communication. Since all the automata are using the $L_{R-I}$ update scheme and because the agents act as selfish reinforcement learners, all the automata at the top level will eventually converge to a pure action and so will at least one of the automata at the lower level. If all the hierarchies are converged to a single path in their hierarchy, the agents are also converged to a unique path in a multi-stage game. The pseudo-code of the exploration phase, which resembles the learning algorithm for hierarchies using the Monte Carlo reward [6], can be seen in Algorithm 1. The difference between the two algorithms is that a separate combined reward must be computed for each agent (since the rewards for all the agents might be different). Another addition is that each automaton keeps track of the global average pay-off received throughout the whole game, as well as the average pay-off during the last exploration phase.

After a given period of time (during which the agents converged, in the algorithm if $s = N$)) the communication phase takes place (see Algorithm 2). The communication phases uses only a few time steps and work as follows: the agents communicate their global average pay-off and the average-pay-off they received

**Algorithm 1.** Pseudo-code for the independent learning phase of the hierarchical periodic policies

```
current phase time step $s \leftarrow 0$
repeat
    $t \leftarrow t + 1$
    $s \leftarrow s + 1$
    for each hierarchy $h \leq H$ do
        Activate the top-level automaton $A_h^0|0$
    end for
    repeat
        for each hierarchy $h \leq H$ do
            Let the active automaton select an action: $a_{hi}$
        end for
        Implement the joint action $(a_{1i}, a_{2j}, \ldots, a_{Hm})$ and observe the immediate reward vector $R$
        for each hierarchy $h \leq H$ do
            Activate the next automaton based on action $a_{hn}$
        end for
    until $l = L$ (the number of levels in the hierarchy)
    A reward is now given to each agent: $r^h$
    for each hierarchy $h \leq H$ do
        for each level $l \leq L$ do
            Update the global average reward: $v^h \leftarrow \frac{v^h(t-1)}{t} + \frac{r^h}{t}$
            Update the average reward of current phase: $w^h \leftarrow \frac{w^h(s-1)}{s} + \frac{r^h}{s}$
            Update the learning automaton that was active during the last trial:
            assume that this automaton chose action $a_{h|i}^l(t)$ during the last episode
            $p_i(t+1) = p_i(t) + \alpha r^h(1 - p_i(t))$ (i.e. update the probability of that action positively)
            $p_j(t+1) = p_j(t) + \alpha r^h p_j(t) \forall j \neq i$ (i.e. update the probability of all other actions negatively)
        end for
    end for
until $s = N$ (the time each independent exploration phase lasts)
```

during the last learning phase to their opponent. The agent that is excluding actions checks whether he still is the best performing agent (i.e. a higher global average and a better average during the last phase). If he still is performing better he excludes another action. The agent starts by excluding the action he converged to at the bottom level. During the next phase, if he is still the best performing agent he will exclude another action at the same level. If at one point, there are no more actions available at this level, the agent will exclude an action of the level above and as such making his way up in the hierarchy. If the agent is not excluding he checks whether he is switched from the worst performing automaton to the better learning automaton. If this is the case it is now his turn to exclude one or more actions.

**Algorithm 2.** Pseudo-code for the exclusion or communication phase of the hierarchical periodic policies

```
for each hierarchy $h \leq H$ do
    broadcast $v^h$ to all $h' \neq h$
    broadcast $w^h$ to all $h' \neq h$
    for each $h' \neq h$ do
        receive from broadcast $v^{h'}$
        receive from broadcast $w^{h'}$
    end for
    if $X =$ true then
        if $v^h < v^{h'}$ AND $w^h > w^{h'} \forall h' \neq h$ then
            $X \leftarrow$ false
        else if $v^h > v^{h'}$ AND $w^h > w^{h'} \forall h' \neq h$ then
            Exclude action at lowest level possible
        else if $v^h > v^{h'}$ AND $w^h > w^{h'} \forall h' \neq h$ then
            $X \leftarrow$ false
            Include all actions
        end if
    else
        if $v^h > v^{h'}$ AND $w^h > w^{h'} \forall h' \neq h$ then
            $X \leftarrow$ true
            Exclude action at lowest level possible
        end if
    end if
    $v^h \leftarrow 0$
end for
```

## 5 Related Work

A previous attempt to solve conflicting multi-stage games using learning automata by Zhou et al. [7] consisted of a 2 level chain of learning automata.

In this setting there are 2 agents both with 2 connected learning automata players. The restriction Zhou et al. put on both games is that they are zero sum games.

The authors showed that the linear chain of learning automata setup asymptotically converges to playing game $B$ if either one of the matrices ($A$ or $B$) has a pure Nash equilibrium and thus one of the two agents always deviates from choosing game $A$, resulting in non-cooperation. Furthermore, the authors showed that if game $B$ has a pure Nash equilibrium, the learning automata at the lower level will converge to this equilibrium. Producing a similar theoretical result for a hierarchy of learning automata is a challenging problem that we do not tackle in this paper.

Instead, we use these results as the baseline for our own experiments. We present empirical results of hierarchical periodic policies and show that they are capable of finding fair solutions in a larger variety of conflicting multi-stage games. In the multi-stage game explained above, the agents at the lower level can only be situated in either game $A$ or game $B$ because the joint-actions at the high level lead to either one of these games. Now consider the problem where each joint-action at the top level leads to a different game: $A$, $B$, $C$ or $D$.

## 6 Experiments

### 6.1 Zero-Sum ABBB

The first game under study is depicted in Table 2. Note that we omit the reward matrix of the first stage, since no rewards are given in this stage. In this game $A$ is only played if both automata at the top level chose to play game $A$, otherwise game $B$ is played. Because these matrices do not have pure Nash equilibria, experimental results have showed that the top automata in a linear chain of automata oscillate between which game to choose and the automata at the lower level oscillate between which strategy to choose. If the same game is played with hierarchical learning automata (Figure 1), there are less oscillations and the automata will eventually converge. However, to which action the automata converge can differ each run.

**Table 2.** Reward matrices of a multi-stage game without pure Nash equilibria

$$A = \begin{pmatrix} 0.6, 0.4 & 0.2, 0.8 \\ 0.35, 0.65 & 0.9, 0.1 \end{pmatrix} B = \begin{pmatrix} 0.4, 0.6 & 0.8, 0.2 \\ 0.65, 0.35 & 0.1, 0.9 \end{pmatrix}$$

A fair and stable solution can be obtained using the hierarchical periodic policies. Figure 2(a) shows the average reward over time for the hierarchical agents with and without periodic policies. The figure shows that the agents without periodic policies obtain an unfair equilibrium in which Agent 1 has a higher reward (*HLA Agent 1* line $\approx 0.58$) than Agent 2 (*HLA Agent 2* line $\approx 0.42$). This result is averaged over 50 runs. By applying the periodic policies, the agents are able to alternate between periods of optimal play for each agent and they find a fair overall solution ($\approx 0.5$).

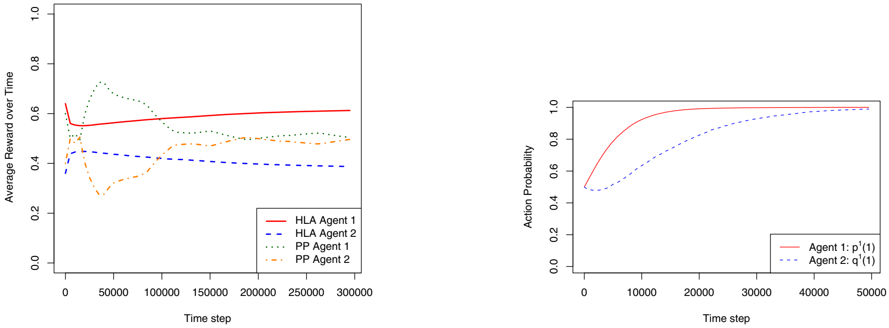
### 6.2 Zero-Sum ABCD

Consider the game with the reward matrices giving in Table 3 in which each joint-action at the top level leads to a different reward matrix at the lower level. In this game all the reward matrices have a pure Nash equilibrium located at position $(1, 1)$. If we let hierarchical agents play this game, the results in Figure 2(b) are obtained. We can see that both agents agree to play game $A$. The reason for this is the following. The combined matrix that is constructed using the four Nash equilibria of all games is shown in Table 4. By analyzing this matrix, we see that there is a pure Nash equilibrium located at position $(1, 1)$ and it is in fact this pure Nash equilibrium to which the agents agree on playing.

This conjecture can be confirmed if we let agents play games in which the matrix of the Nash equilibria does not contain a Nash equilibrium itself, the agents show oscillating behavior.

**Table 3.** Reward matrices

$$A = \begin{pmatrix} 0.6, 0.4 & 0.8, 0.2 \\ 0.35, 0.65 & 0.9, 0.1 \end{pmatrix} \; B = \begin{pmatrix} 0.7, 0.3 & 0.9, 0.1 \\ 0.6, 0.4 & 0.8, 0.2 \end{pmatrix}$$
$$C = \begin{pmatrix} 0.5, 0.5 & 0.7, 0.3 \\ 0.3, 0.7 & 0.8, 0.2 \end{pmatrix} \; D = \begin{pmatrix} 0.3, 0.7 & 0.7, 0.3 \\ 0.1, 0.9 & 0.2, 0.8 \end{pmatrix}$$



(a) Average rewards for hierarchical agents with (*PP Agent 1* and *PP Agent 2*) and without (*HLA Agent 1* and *HLA Agent 2*) periodic policies playing the game in Table 2 with $\alpha = 0.005$ and $s = 37500$.

(b) The probability to select game $A$ for the top learning automata of both agents ($\alpha = 0.005$).

**Fig. 2.**

In Table 5 is such a game depicted. Figure 3(a)-3(b) shows the evolution of the action probabilities of the automata of the top level with and without periodic polices. The predicted oscillations are clearly present for the hierarchical learning automata. The action probabilities of the hierarchical periodic policies seem to oscillate, yet this is due to the exclusion of the actions. The average reward over time (Figure 4) clearly show that the agents find a fair solution.

### 6.3 Non-zero Sum Games

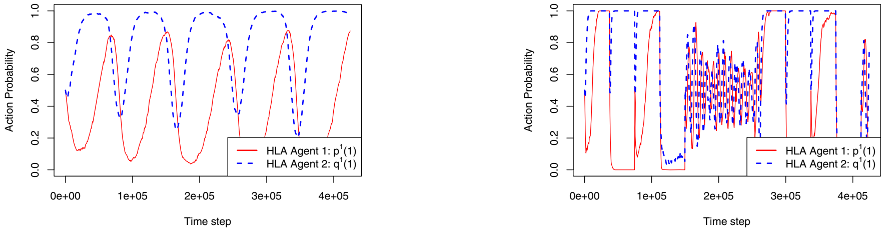
While the theory of Zhou et al. was limited to games of a certain structure as was mentioned above, a second limitation was that the matrices had to be zero-sum games. The periodic policies technique is created to function in general

**Table 4.** Upper left: the Nash equilibrium of game $A$. Upper right: the Nash equilibrium of game $B$. Lower left: the Nash equilibrium of game $C$. Lower right: the Nash equilibrium of game $D$.

$$M^{\text{Nash equilibria}} = \begin{pmatrix} 0.6, 0.4 & 0.7, 0.3 \\ 0.5, 0.5 & 0.3, 0.7 \end{pmatrix}$$

**Table 5.** Conflicting multi-stage game in which each joint-action at the top level leads to a different reward matrix

$$A = \begin{pmatrix} 0.6, 0.4 & 0.8, 0.2 \\ 0.35, 0.65 & 0.9, 0.1 \end{pmatrix} \quad B = \begin{pmatrix} 0.3, 0.7 & 0.7, 0.3 \\ 0.1, 0.9 & 0.2, 0.8 \end{pmatrix}$$
$$C = \begin{pmatrix} 0.5, 0.5 & 0.7, 0.3 \\ 0.3, 0.7 & 0.8, 0.2 \end{pmatrix} \quad D = \begin{pmatrix} 0.7, 0.3 & 0.9, 0.1 \\ 0.6, 0.4 & 0.8, 0.2 \end{pmatrix}$$



(a) The probability to select game $A$ for the top learning automata of both agents ($\alpha = 0.001$).

(b) The probability to select game $A$ for the top learning automata of both agents using periodic policies ($\alpha = 0.001$).

**Fig. 3.**

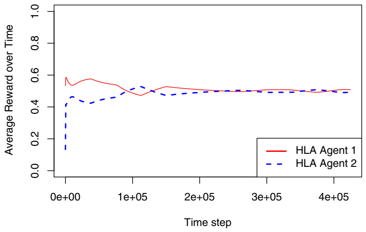


**Fig. 4.** Average reward over time for the hierarchical periodic policies playing the game in Table 5

sum conflicting interest games. Consider the game with reward matrices of Table 6. The optimal reward an agent can receive, without considering the reward of the other agent, has a value of 1.0. The fairest solution of the game would be to alternate between the two pure Nash equilibria of the Battle of the Sexes. This would give an average reward of 0.75 to both agents. Figure 5(a) shows the average reward over time for the hierarchical agents with periodic policies. The result shows that the agents obtain a fair solution that is averaged around 0.73 which is very close to the theoretical optimal fair solution.

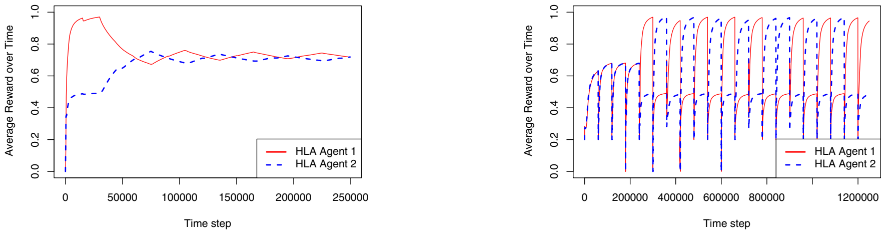
In a final experiment, the reward matrices are mixed conflicting and common interest games (see Table 7). The optimal strategy for the agents would be to alternate between the pure Nash equilibria of the Battle of the Sexes game, giving an average reward of 0.75. However, there is also a common interest game leading

**Table 6.** Reward matrices of a general sum game. The matrices $A$ and $D$ are stochastic versions of the Prisoner's Dilemma and matrices $B$ and $C$ are stochastic version of the Battle of the Sexes.

$$A = \begin{pmatrix} 0.2, 0.2 & 1.0, 0.0 \\ 0.0, 1.0 & 0.4, 0.4 \end{pmatrix} \quad B = \begin{pmatrix} 1.0, 0.5 & 0.0, 0.0 \\ 0.0, 0.0 & 0.5, 1.0 \end{pmatrix}$$
$$C = B \qquad\qquad D = A$$

**Table 7.** Reward matrices of a general sum game. The matrices $A$ is the stochastic version of the Prisoner's Dilemma and matrices $B$, $C$ and $D$ are common interest games.

$$A = \begin{pmatrix} 1.0, 0.5 & 0.0, 0.0 \\ 0.0, 0.0 & 0.5, 1.0 \end{pmatrix} \quad B = \begin{pmatrix} 0.2, 0.2 & 0.2, 0.2 \\ 0.2, 0.2 & 0.2, 0.2 \end{pmatrix}$$
$$C = B \qquad\qquad D = \begin{pmatrix} 0.7, 0.7 & 0.3, 0.3 \\ 0.3, 0.3 & 0.7, 0.7 \end{pmatrix}$$



(a) Average reward over time for the hierarchical periodic policies playing the game in Table 6 ($\alpha = 0.002$).

(b) Average reward of each independent learning phase for the hierarchical periodic policies playing the game in Table 7 ($\alpha = 0.002$).

**Fig. 5.**

to a pay-off of 0.7 for both agents. Figure 5(b) shows the average reward of each independent learning phase of the hierarchical agents with periodic policies. We can see that the agents alternate between the two pure Nash equilibria of the Battle of the Sexes game. The same experiment without the use of periodic policies gives an average reward of 0.66 (averaged over 100 runs).

## 7 Conclusion

In this paper we enhanced independent agents with social behavior based on group characteristics of the homo egualis society. We defined a fair as a solution concept where the relative difference between the agents is minimized. By extending hierarchical learning automata with the ability to (temporarily) exclude actions, they were able to find different periodic policies which favored either one of the players. The general result is that we were able to create a distributed framework that can find a fair solution in pure conflicting games, in particular zero-sum games. The method we described in this paper is suited for finding fair solutions in conflicting zero-sum and general-sum games that are not bound to a certain structure (which was the case for the linear chain of learning automata).

In systems where the global performance is only as good as the performance of the worst performing agent this solution concept of periodic policies can be used. Examples of applications that fall under this category are load balancing, job scheduling or network routing. The idea of alternating between policies by temporarily excluding actions originated in single-stage games where it proved its value in a single-stage job scheduling application [8]. In future research, periodic policies will be applied to games with reward matrices at all the different stages of the sequential decision problem.

## References

1. Fehr, E., Schmidt, K.M.: A theory of fairness, competition and cooperation. Quarterly Journal of Economics 114, 817–868 (1999)
2. Verbeeck, K., Nowé, A., Parent, J., Tuyls, K.: Exploring selfish reinforcement learning in repeated games with stochastic rewards. Journal of Autonomous Agents and Multi-agent Systems 14(3), 239–269 (2006)
3. Verbeeck, K.: Coordinated Exploration in Multi-Agent Reinforcement Learning. PhD thesis, Vrije Universiteit Brussel (2004)
4. Narendra, K.S., Thathachar, M.A.L.: Learning Automata: An Introduction. Prentice Hall, Englewoord Cliffs (1989)
5. Nowé, A., Verbeeck, K.: Distributed reinforcement learning: Loadbased routing a case study. In: Notes of the Neural, Symbolic and Reinforcement Methods for Sequence Learning Workshop, Stockholm, Sweden (1999)
6. Narendra, K.S., Parthasarathy, K.: Learning automata approach to hierarchical multiobjective analysis. IEEE Transactions on Systems, Man, and Cybernetics 21(2), 263–273 (1991)
7. Zhou, J., Billard, E., Lakshmivarahan, S.: Learning in multi-level games with incomplete information - part ii. IEEE Transactions on Systems, Man, and Cybernetics Part B 29(3), 340–399 (1999)
8. Verbeeck, K., Nowé, A., Parent, J.: Homo egualis reinforcement learning agents for load balancing. LNCS (LNAI), vol. 2564, pp. 81–91. Springer, Heidelberg (2002)

# Incremental Learning Method of Simple-PCA

Tadahiro Oyama[1], Stephen Karungaru[1], Satoru Tsuge[1],
Yasue Mitsukura[2], and Minoru Fukumi[1]

[1] University of Tokushima, Department of Information Science and Intelligent
Systems, 2-1, Minami-Josanjima, Tokushima, 770-8506, Japan
{oyama,karunga,tsuge,fukumi}@is.tokushima-u.ac.jp
[2] Tokyo University of Agriculture and Technology, Graduate School of
Bio-Applications and Systems Engineering, 2-24-16, Naka-cho, Higashi-koganei,
Tokyo, 184-8588, Japan
mitsu_e@cc.tuat.ac.jp

**Abstract.** In this paper, we propose an incremental learning algorithm named Incremental Simple-PCA. This algorithm is added an incremental learning function to the Simple-PCA that is an approximation algorithm of the principal component analysis where an eigenvector can be calculated by a simple repeated calculation. Using the proposed algorithm, it allows to update faster the eigenvector by using incremental data. To verify the effectiveness of this algorithm, we carry out computer simulations on personal authentication that uses face images and wrist motion discrimination using wrist EMG by incremental learning. As a result, we can confirm the effectiveness from the aspects of accuracy and a computing time by comparing the Incremental PCA that gave the incremental learning function to the conventional PCA.

**Keywords:** PCA, Simple-PCA, Incremental PCA, Incremental Simple-PCA, Incremental learning, face recognition.

## 1 Introduction

In a field of pattern recognition, researches on feature extraction and dimension reduction are actively done by using methods of the principal component analysis (PCA) and the linear discriminant analysis (LDA), etc. to data with a lot of attributes. Especially, the effectiveness using PCA by a research on facial images has been demonstrated [1,2,3]. Moreover, Incremental PCA (IPCA) exists as an algorithm that added an incremental learning function to PCA [4,5,6,7]. By using this algorithm, even if a learning data is newly added, it enables us to learn it sequentially. There are two advantages in this algorithm. First, this algorithm has the advantage that it is not necessary to restart learning from the beginning. Moreover, the learning data need not be preserved. However, a calculation cost becomes huge, because PCA and IPCA need a matrix calculation. When we build in PCA and IPCA into a device that should operate in real-time, the calculation amount especially becomes a serious problem.

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 403–410, 2008.

There is an algorithm named Simple-PCA to solve such a problem. This is an approximation algorithm of PCA where the principal component vector can be calculated by simple repeated calculation. This algorithm is used in many fields such as hand-written characters and dimensionality reduction of a model [8,9]. However, the algorithm that added the learning function to this Simple-PCA does not exist.

In this paper, we propose an Incremental Simple-PCA that added an incremental learning function to the Simple-PCA. Moreover, we carry out the computer simulation on personal authentications that use face images and wrist motion discrimination using wrist EMG by incremental learning to verify the effectiveness of the proposal algorithm.

## 2 PCA and Incremental PCA

PCA exists as a method of the dimension reduction and the feature extraction to high-dimensional data with a lot of features. Moreover, there is Incremental PCA (IPCA) as the algorithm to which the eigenvector can be updated by incrementing the new data to the PCA. In this section, we describe these algorithms.

### 2.1 PCA

In this section, we explain PCA. First, a set of p-dimensional data $\boldsymbol{x}_i$ to use is defined as $\boldsymbol{x} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\}$, where $n$ is the number of data. In this case, the eigenvector can be obtained by solving the eigenvalue problem of covariance matrix $\boldsymbol{C}$ shown in eq. (1).

$$\boldsymbol{C} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x_i} - \bar{\boldsymbol{x}})(\boldsymbol{x_i} - \bar{\boldsymbol{x}})^T \tag{1}$$

where mean vector $\bar{\boldsymbol{x}}$ is shown as follows.

$$\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i \tag{2}$$

The feature vector $\boldsymbol{a}_i$ is obtained by projecting the vector $\boldsymbol{x}_i$ to the eigenvector $\boldsymbol{U}$ derived from this eigenvalue problem.

### 2.2 Incremental PCA

Incremental PCA (IPCA) is the algorithm that P.M.Hall et al. proposed [4], and it is used for the localization control of the mobile robot [5,7] and the image processing in online [6]. We describe detail of this algorithm in this section.

First of all, we assume that we have already built the eigenspace (eigenvector) $\boldsymbol{U}$=$\boldsymbol{u}_j(j = 1...p)$ by using the data of $\boldsymbol{x}_i(i = 1, 2, ..., n)$. Moreover, the mean vector is $\bar{\boldsymbol{x}}$, the corresponding eigenvalues are $\boldsymbol{\lambda} = diag(\boldsymbol{\Lambda})$, the new input data is assumed to $\boldsymbol{x}_{n+1}$.

To update the eigenvector, the mean vector $\bar{\boldsymbol{x}}$ is updated by using the new data $\boldsymbol{x}_{n+1}$.

$$\bar{\boldsymbol{x}}' = \frac{1}{n+1}(n\bar{\boldsymbol{x}} + \boldsymbol{x}_{\boldsymbol{n+1}}) \tag{3}$$

Next, the new data $\boldsymbol{x}_{n+1}$ is projected to the current eigenvector $\boldsymbol{U}$, and the feature vector $\boldsymbol{a}_{n+1}$ is obtained. This is the vector that represents the new data in the current eigenspace.

$$\boldsymbol{a}_{n+1} = \boldsymbol{U}^T(\boldsymbol{x}_{n+1} - \bar{\boldsymbol{x}}) \tag{4}$$

The residual vector $\boldsymbol{h}_{n+1}$ is calculated by using this feature vector $\boldsymbol{a}_{n+1}$. This residual vector is orthogonal to the eigenvector.

$$\boldsymbol{h}_{n+1} = (\boldsymbol{U}\boldsymbol{a}_{n+1} + \bar{\boldsymbol{x}}) - \boldsymbol{x}_{n+1} \tag{5}$$

We normalized $\boldsymbol{h}_{n+1}$ that is necessary to update the eigenvector according to the condition in eq. (6).

$$\hat{\boldsymbol{h}}_{n+1} = \begin{cases} \frac{\boldsymbol{h}_{n+1}}{\|\boldsymbol{h}_{n+1}\|_2} & if \parallel \boldsymbol{h}_{n+1} \parallel_2 \neq 0 \\ \boldsymbol{0} & otherwise \end{cases} \tag{6}$$

The updated new eigenvector $\boldsymbol{U}'$ is obtained by appending $\hat{\boldsymbol{h}}_{n+1}$ to the current eigenvector $\boldsymbol{U}$ and rotating them using rotation matrix $\boldsymbol{R}$.

$$\boldsymbol{U}' = [\boldsymbol{U}, \hat{\boldsymbol{h}}_{n+1}]\boldsymbol{R} \tag{7}$$

$\boldsymbol{R}$ is derived by solving the eigenvalue problem shown in following eq. (8).

$$\boldsymbol{D}\boldsymbol{R} = \boldsymbol{R}\boldsymbol{\Lambda}' \tag{8}$$

where matrix $\boldsymbol{D}$ is composed as follows.

$$\boldsymbol{D} = \frac{1}{n+1}\begin{bmatrix} \boldsymbol{\lambda} & \boldsymbol{0} \\ \boldsymbol{0}^T & 0 \end{bmatrix} + \frac{n}{(n+1)^2}\begin{bmatrix} \boldsymbol{a}\boldsymbol{a}^T & \gamma\boldsymbol{a} \\ \gamma\boldsymbol{a}^T & \gamma^2 \end{bmatrix} \tag{9}$$

where $\gamma$ is obtained by performing calculation shown in eq. (10).

$$\gamma = \hat{\boldsymbol{h}}_{n+1}(\boldsymbol{x}_{n+1} - \bar{\boldsymbol{x}}) \tag{10}$$

We can obtain the updated feature vector by projecting data to new eigenvector $\boldsymbol{U}'$ calculated in eq. (7).

## 3 Simple-PCA and Incremental Simple-PCA

The Simple-PCA is an algorithm that can calculate an approximation vector of the principal component vector by the simple repeated calculation. In this section, we explain the algorithm of this Simple-PCA. Moreover, the algorithm of Incremental Simple-PCA that gives an incremental learning function to Simple-PCA is explained.

### 3.1 Simple-PCA

The Simple-PCA (Simple Principal Component Analysis) is a technique which is proposed by Partridge and others [8] to speed up the principal component analysis. The technique is the approximation algorithm from which principal components can be sequentially found from the first component. Moreover, its effectivity is confirmed in many fields such as recognition of hand-written characters, dimensionality reduction of a model for information retrieval and recognition of using face images and so on [8,9,10,11].

The algorithm of this technique sequentially solves for eigenvectors that maximizes the variance over all samples. Concretely, it is summarized as follows.

First of all, a set of vectors to use is defined as follows.

$$\boldsymbol{v} = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_m\} \tag{11}$$

To make the center of gravity of this set the origin the calculation shown in eq. (13) is performed, and a new set of vectors (12) is obtained.

$$\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_m\} \tag{12}$$

$$\boldsymbol{x}_i = \boldsymbol{v}_i - \frac{1}{m}\sum_{j=1}^{m} \boldsymbol{v}_j \tag{13}$$

Next, the following output function is used.

$$y_n = (\boldsymbol{\alpha}_n)^T \boldsymbol{x}_j \tag{14}$$

where $\boldsymbol{\alpha}_n$ is a eigenvector that shows the n-th principal component. If the input vector $\boldsymbol{x}_j$ is the same direction as $\boldsymbol{\alpha}_n$, the function shown by eq. (14) outputs a positive value. If it is the opposite direction, a negative value is used. Thus, the following threshold function is introduced.

$$\boldsymbol{f}(y_n, \boldsymbol{x}_j) = \begin{cases} \boldsymbol{x}_j & if \quad y_n \geq 0 \\ -\boldsymbol{x}_j & otherwise \end{cases} \tag{15}$$

The initial vector $\boldsymbol{\alpha}_n^1$ initialized by arbitrary random values is brought close in the same direction as $\boldsymbol{\alpha}_n$ by these functions and the repetition operation shown in eq. (16).

$$\boldsymbol{\alpha}_n^{k+1} = \frac{\sum_j \boldsymbol{f}(y_n, \boldsymbol{x}_j)}{\parallel \sum_j \boldsymbol{f}(y_n, \boldsymbol{x}_j) \parallel}, y_n = (\boldsymbol{\alpha}_n^k)^T \boldsymbol{x}_j \tag{16}$$

where $k$ is number of repetitions. Moreover, $\boldsymbol{\alpha}_n^{k+1}$ is a vector after calculating $k+1$ times. The value of the output function is calculated by using $\boldsymbol{\alpha}_n^k$ that is the previous calculation result. Furthermore, this repetition calculation is done until $\boldsymbol{\alpha}_n^{k+1}$ is converged. This vector after it converges is an eigenvector.

When the next eigenvector is calculated, it is necessary to calculate it by using a new vector $\boldsymbol{x}'_j$ after the previous principal component is removed from the input vector by doing the calculation shown in eq. (17).

$$\boldsymbol{x}'_j = \boldsymbol{x}_j - (\boldsymbol{\alpha}_n^{k+1} \cdot \boldsymbol{x}_j)\boldsymbol{\alpha}_n^{k+1} \tag{17}$$

After the component is removed, the principal component can be evaluated by repeating a similar calculation in order with a high accumulated relevance.

### 3.2 Incremental Simple-PCA

The improvement algorithm of Simple-PCA aimed to update the eigenvector (principal component vector) according to incremental data. First of all, incremental data is defined as $\boldsymbol{v}_{m+1}$ and the last mean value is assumed to be $\bar{\boldsymbol{v}}$. Moreover, $m$ is the number of data until last time. The calculation shown in eq. (18) is executed, a new mean $\bar{\boldsymbol{v}}'$ is found.

$$\bar{\boldsymbol{v}}' = \frac{1}{m+1}(m\bar{\boldsymbol{v}} + \boldsymbol{v}_{m+1}) \tag{18}$$

$\boldsymbol{x}_{m+1}$ is obtained by using this $\bar{\boldsymbol{v}}'$.

$$\boldsymbol{x}_{m+1} = \boldsymbol{v}_{m+1} - \bar{\boldsymbol{v}}' \tag{19}$$

Next, we introduce the following threshold functions as well as the case of Simple-PCA.

$$y_n = (\boldsymbol{\alpha}_n)^T \boldsymbol{x}_{m+1} \tag{20}$$

$$\boldsymbol{f}(y_n, \boldsymbol{x}_{m+1}) = \begin{cases} \boldsymbol{x}_{m+1} & if \quad y_n \geq 0 \\ -\boldsymbol{x}_{m+1} & otherwise \end{cases} \tag{21}$$

where $\boldsymbol{\alpha}_n$ shows a n-th eigenvector calculated last time. The new eigenvector $\boldsymbol{\alpha}'_n$ is obtained by calculating the following eq. (22)

$$\boldsymbol{\alpha}'_n = \frac{1}{m+1}\left(\frac{\boldsymbol{f}(y_n, \boldsymbol{x}_{m+1})}{\|\boldsymbol{f}(y_n, \boldsymbol{x}_{m+1})\|}\right) + \frac{m}{m+1}\boldsymbol{\alpha}_n \tag{22}$$

Finally, the previous principal component $\boldsymbol{\alpha}'_n$ is removed from the new data $\boldsymbol{x}_{m+1}$ by the eq. (23).

$$\boldsymbol{x}'_{m+1} = \boldsymbol{x}_{m+1} - (\boldsymbol{\alpha}'_{\boldsymbol{n}} \cdot \boldsymbol{x}_{m+1})\boldsymbol{\alpha}'_{\boldsymbol{n}} \tag{23}$$

The eigenvector is sequentially updated by repeating these calculations. However, this algorithm is an approximation algorithm, and there is a problem that the previous principal component has not been removed from the initial data.

## 4 Incremental Learning Experiment

To verify the effectiveness of Incremental Simple-PCA that we had proposed, the personal authentication experiment that used face images and wrist motion recognition experiment that used wrist EMG signal were carried out. At the same time, we conducted the experiment that used Incremental PCA, and compared these results from the aspects of the recognition accuracy and the computing time.

### 4.1 Personal Authentication Experiment

In this paper, we performed the personal authentication by using face images as the incremental learning experiment. In the incremental learning, the new person's image is used as the incremental data. The new person indicates the person that has not been registered in the database beforehand. In this paper, we used the face images of the database of the University of Oulu. The images are cut out the part of the face, reduced to 25 25 dimensions, and used as input data. The outline of the experiment is shown in Fig.1. 5 images per person that are taken in the different lighting conditions are prepared for 20 people (100 images in total). PCA and Simple-PCA are performed to these images. In consequence, we obtain the eigenvectors (eigenface) in the initial state. The personal authentication is carried out by the nearest neighbour method in the eigenspace created by using the obtained eigenvectors. An image per person is used for the evaluation data. Furthermore, the leave-one-out cross-validation method is used as the evaluation method. The recognition result in the initial state is obtained under the above-mentioned conditions.

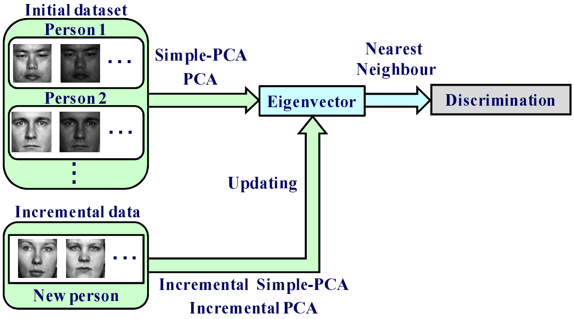


**Fig. 1.** Outline of experiment

Next, we increment images of the new person at random as incremental learning data. The incremental data used is the new 10 person images (5 images per person). The eigenvectors are updated by using Incremental PCA and Incremental Simple-PCA whenever data is incremented. As a comparative experiment, we tried the incremental learning experiment by using the eigenvectors that are not updated with the incremental data, which means eigenvectors in the initial state.

The incremental learning experimental result is shown in Fig.2. The horizontal axis is the number of incremental data, the vertical axis is recognition accuracy. The recognition accuracy in the initial state is about 78% in PCA and about 83% in Simple-PCA respectively. In the initial state, when Simple-PCA is used, it is found that the recognition accuracy is high compared with the case when PCA is used.

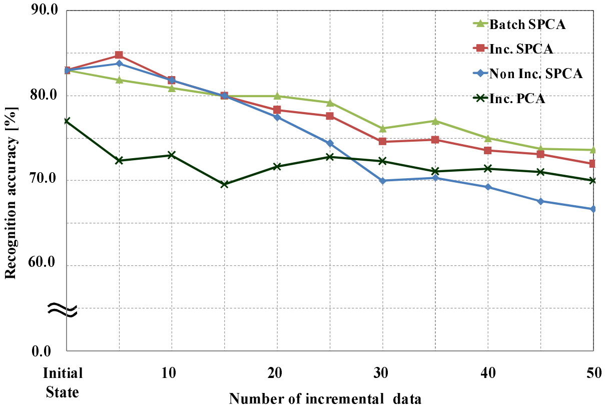


**Fig. 2.** Recognition accuracy in the experiment

In the incremental learning, when the eigenvector is updated by using Incremental Simple-PCA, the recognition accuracy is higher than the case when the eigenvector is not updated. Moreover, the recognition accuracy is high compared with the case when Incremental PCA is used. Therefore, it is thought that Incremental Simple-PCA is the effective algorithm in incremental learning.

### 4.2 Wrist Motion Recognition Experiment

In addition to the personal authentication experiment, the wrist motion recognition experiment that used EMG is carried out. As a result, we can confirm the effectiveness of proposed algorithm as well as the result of the personal authentication experiment that used face images.

### 4.3 Computational Time

In this subsection, we compare the computational time of Incremental PCA and Incremental Simple-PCA. Each result is shown in Table 1. The computational time was measured every five incremental data. In Incremental PCA, the computational time is very long, and time has been extended as the number of incremental data increases. On the other hand, the computational time of Incremental Simple-PCA is very short. As a result, it can be said that this algorithm is efficient when real-time performance is needed.

**Table 1.** Computational time of each method

| No. of incremental data | Initial | 5 | 10 | 15 | 20 | 25 | 30 |
|---|---|---|---|---|---|---|---|
| Incremental PCA[sec] | 3977.180 | 0.703 | 1.469 | 2.469 | 3.508 | 4.945 | 7.086 |
| Incremental Simple-PCA[sec] | 0.216 | 0.016 | 0.016 | 0.015 | 0.016 | 0.015 | 0.016 |

## 5 Conclusion

In this paper, we proposed Incremental Simple-PCA that equipped the incremental learning function to Simple-PCA, which was the approximation algorithm of the principal component analysis (PCA). The incremental learning was tried by applying this algorithm to the recognition experiments that used the face images and EMG signal. As a result, the availability of the proposed algorithm was able to be confirmed from the aspects of the accuracy and the computational time.

In the future, we will build in this algorithm into the device that should operate in real-time.

## References

1. Akamatsu, S.: Computer Recognition of Human Face -A Survey. IEICE Trans. D J80-D2(8), 2031–2046 (1998) (in Japanese)
2. Turk, M., Pentland, A.: Eigenfaces for recognition. Journal of Cognitive Neuroscience 3(1), 71–86 (1991)
3. Pentland, A., Moghaddam, B., Starner, T.: View-based and modulear eigenspaces for face recognition. In: Proc. of CVPR 1994, pp. 84–91 (1994)
4. Hall, P.M., Marshall, D., Martin, R.R.: Incremental Eigenanalysis for Classification. In: Proc. of the British Machine Vision Conference, vol. 1, pp. 286–295 (1998)
5. Artac, M., Jogan, M., Leonardis, A.: Mobile robot localization using an incremental eigenspace model. In: Proc. of IEEE International Conference on Robotics and Automation, Washington,D.C, pp. 1025–1030 (2002)
6. Artac, M., Jogan, M., Leonardis, A.: Incremental PCA for On-line Visual Learning and Recognition. In: Proceedings of the 16th International Conference on Pattern Recognition(ICPR), Quebec City, Canada, pp. 781–784 (2002)
7. Freitas, R., Santos-Victor, J., Sarcinelli-Filho, M., Bastos-Filho, T.: Performance Evaluation of Incremental Eigenspace Models for Mobile Robot Localization. In: Proceedings of the IEEE 11th International Conference on Advanced Robotics (ICAR 2003), Coimbra, Portugal, pp. 417–422 (2003)
8. Partridge, M., Calvo, R.: Fast dimentionality reduction and simple PCA. In: IDA, vol. 2, pp. 292–298 (1997)
9. Kuroiwa, S., Tsuge, S., Tani, H., Tai, X.-Y., Shishibori, M., Kita, K.: Dimensionality reduction of vector space model based on Simple PCA. In: Proc. Knowledge-Based Intelligent Information Engineering Systems & Allied Technologies (KES), Osaka, vol. 2, pp. 362–366 (2001)
10. Nakano, M., Yasukata, F., Fukumi, M.: Recognition of Smiling Faces Using Neural Networks and SPCA. International Journal of Computational Intelligence and Applications 4(2), 153–164 (2004)
11. Takimoto, H., Mitsukura, Y., Fukumi, M., Akamatsu, N.: A Feature Extraction Method for Personal Identification System by Using Real-Coded Genetic Algorithm. In: Proc. of 7th SCI 2003, Orlando, USA, vol. 4, pp. 66–77 (2003)

# Utilization of Soft Computing Techniques for Making Environmental Games More Exciting –Toward an Effective Utilization of the COMMONS GAME

Norio Baba[1], Kenta Nagasawa[1], and Hisashi Handa[2]

[1] Information Science, Osaka Kyoiku University,
Asahiga-Oka, 4-698-1, Kashiwara City, Osaka Prefecture, 582-8582, Japan
[2] Information Science, Okayama University, Tsushima Naka, 3-1-1, Okayama City,
700-8530, Japan

**Abstract.** We have suggested that utilization of soft computing techniques such as GAs and EAs could contribute a lot for making the original COMMONS GAME much more exciting. In this paper, we try to find an answer concerning to the question "Which game is the best for letting people consider seriously about the commons among the three games (the original COMMONS GAME, the modified COMMONS GAME utilizing GAs & NNs , and the modified COMMONS GAME utilizing EAs & NNs) ?". Several game playing by our students confirm that the modified COMMONS GAME utilizing EAs & NNs can provide the best chance for letting players consider seriously about the commons.

**Keywords:** Soft computing techniques, Gaming, COMMONS GAME, Player's involvement in game playing, Exciting game.

## 1 Introduction

In recent years, gaming has gradually been recognized by many people as a new and promising tool to deal with complex problems in which human decisions have far reaching effects on others. It has been used for various purposes such as education, training, decision-making, entertainment, and etc.[1]-[8]. Along with the appearance of various types of games, continuous effort has been done in order to let existing games be more exciting [1]-[13]. About a decade ago, we suggested that GAs & NNs could be utilized for making the original COMMONS GAME [4], one of the most popular environmental games, become much more exciting [9], [10]. Recently, we also suggested that EAs & NNs could also be utilized for constructing a more exciting version of the original COMMONS GAME [13].

In this paper, we shall compare those three games (Original COMMONS GAME, COMMONS GAME modified by GAs & NNs, COMMONS GAME modified by EAs & NNs) by checking the data having been obtained after the several game playing of our university (Osaka Kyoiku Univ.) students.

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 411–417, 2008.

## 2 COMMONS GAME

The COMMONS GAME [4], one of the most familiar games dealing with serious environmental issues in this planet, was developed by Powers et al. around 1980. Since we live in a world having only finite natural resources such as pure water, fishes, and forests (commons), it is wise to consider their careful utilization. The COMMONS GAME may be quite helpful in stimulating discussion of this problem. Fig. 1 illustrates the layout of the original COMMONS GAME.

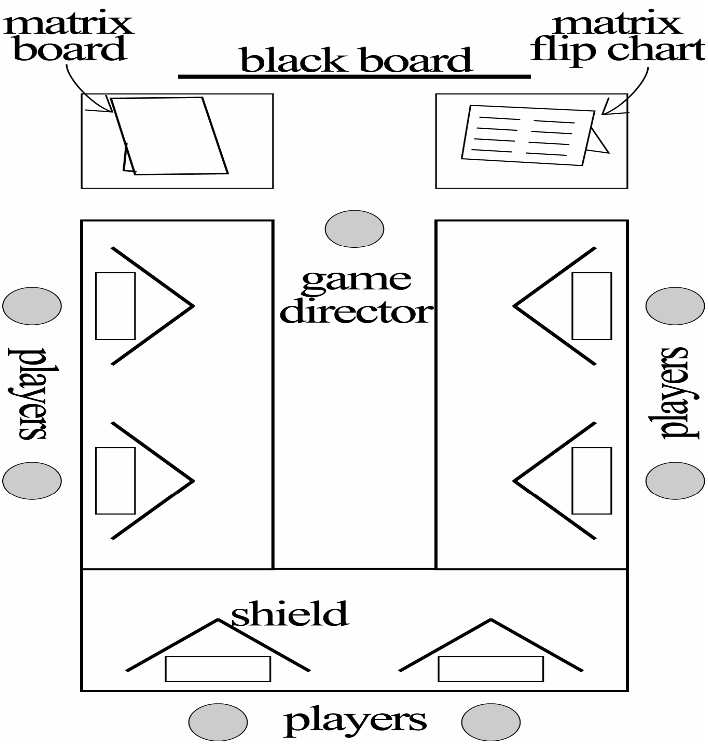


**Fig. 1.** Layout of the original COMMONS GAME

In the following, we give a brief introduction to this game:

First, as shown in Fig.1, six players are asked to sit around a table. Following a brief introduction concerning game playing, the game director informs the players that their objective is to increase their gains by choosing one card among the five colored (Green, Red, Black, Orange, Yellow) cards in each round. In each round, players hide their cards behind a cardboard shield to ensure individual privacy. Each colored card has its own special meaning concerning the attitude toward the environmental protection, and has the following effect upon the total gains of each player:

1) Playing a green card implies selfish exploitation of the commons. Players who play a green card can get a maximum reward. However, they lose 20 points if one of the other players plays a black card in the same round.
2) A red card represents a careful utilization of the commons. Red card players can get about forty percent as much in points that green players have received.
3) A black card has a punishing effect on the green card players. Players who have played a black card have to lose 6 points divided by the number of black card players, but are able to punish green card players by giving them - 20 points.
4) A yellow card implies a complete abstention from utilization of the commons. Players who have chosen this card get 6 points.
5) Orange cards give an encouraging effect to red card players. Players who have chosen this card have to lose 6 points divided by the number of the orange cards in the same round, but are able to add 10 points to red card players.

Depending upon the players' attitude toward environmental protection and their monetary desire, the state of the commons changes. If players are too eager to exploit the commons, then they would face serious deterioration of the commons in a rather early stage of game playing. Although players are informed that there will be 60 rounds, each game ends after 50 rounds. After each 8th round, players have a three minute conference. They could discuss everything about the game and decide every possible way to play in future rounds.

We have so far explained a brief outline of the COMMONS GAME. Due to page, we don't go into further details. (Interested readers are referred to [4], [7], and [8]).

## 3 Modified COMMONS GAME Utilizing Soft Computing Techniques

We have so far enjoyed a large number of playing of the original COMMONS GAME. Those experiences have given us a valuable chance to consider seriously about the current situation of the commons. However, we did often find that some players lost their interest in game playing, even in the middle of the game.

We have tried to find the reason why some players lost their interest in game playing. We have come to the conclusion that the original COMMONS GAME is comparatively monotonous. Further, we have concluded that the following rule makes its game playing monotonous:

In the original COMMONS GAME, players who have chosen a green card receive a penalty, - 20 points, when some player chooses a black card in the same round. On the other hand, black card players receive a minus point [- 6 / (numbers of players who have chosen a black card) ]even though they have contributed a lot in giving punishing effect toward green card players and maintaining current state concerning the commons. Orange card players (who have played an important role in recommending other players use of the red card (not green card)) also receive a minus point [- 6 / (numbers of players who have chosen an orange card) ].

Only red card players who have not played any positive role in maintaining current state of the commons always receive gains.

We have considered that some change in the points "-20" and "-6" mentioned above would make the original COMMONS GAME much more exciting. In order to find an appropriate point for each card, we tried to utilize GAs[10]. Recently, we also tried to utilize EAs [13].

In the following subsections, we shall briefly explain "How the GAs and EAs have been utilized in order to find a better point of each card ?"

### 3.1 Modified COMMONS GAME Constructed by the Use of GAs

We have considered that some changes in the points – 20 and – 6 mentioned above would make game playing of the original COMMONS GAME much more exciting.
In order to find an appropriate point of each card, we have tried to utilize GAs as follows [10]:

1. Each chromosome consists of a 15 bit 0-1 sequence whose first 5 bits represent a point in the closed interval [- 50, - 18]that green players receive when some of the players select a black card, second 5 bits represent a total point in the closed interval [- 6, 26]that black players receive when they select black cards, and last 5 bits represent a total point in the closed interval [- 6, 26] that orange players receive when they select orange cards.
2. First, 30 initial population was chosen randomly. Then, the roulette strategy plus elitist strategy was utilized in order to produce a population in the next generation.
3. In order to construct a fitness function for evaluating each chromosome, we have taken the followings having been observed during each game playing into consideration.
   a) How often a top player has been replaced by other five players ?
   b) Variance of each player's total point. (High fitness value should be given when the variance of each player's total point is low.)
   c) Total number of the black cards having been chosen in the game playing.
   d) Final state of the COMMONS.
   e) Total number of the orange cards having been chosen in the game playing.

After 100 generations, we have obtained the following rule:

1) Black Card: +24 2) Penalty Point of the Green Card: - 42
3) Orange Card: + 14

### 3.2 Modified COMMONS GAME Constructed by the Use of the Two Evolutionary Algorithms

In the original COMMONS GAME, point of each colored card is fixed and any environmental change is not taken into account for deciding it. Recently, we have tried to consider a new rule for assigning a point to each colored card which takes environmental changes into consideration.

In the followings, we shall briefly explain our new trial.
First, we set up the following framework for deciding a point of each colored card:

a) Penalty $P_G$ for green players: We proposed an appropriate way for penalizing green players which takes the environmental changes into account:

$P_G$ = - $W_G$ × (Gain G), where $W_G$ means the numerical value that is determined by the Evolutionary Programming, and "Gain G" denotes the return that the green players can get if any other player does not choose "Black Card."

b) Point AOB that black players lose: We proposed an appropriate way (for asking black players pay cost in trying to penalize green players) which takes the environmental changes into account: AOB = OB / NOB (OB = − A × (Gain R)), where A means the numerical value that is determined by the Evolutionary Programming, and NOB means the number of players who have chosen the black cards, and "Gain R" denotes the return that the red player can get.

c) Point OR that orange players add to the red players: We proposed an appropriate way (for helping red players maintain the commons) which takes the environmental changes into account: OR = $W_0$ × (Gain R), where $W_0$ means the numerical value that is determined by the Evolutionary Programming.

Then, we tried to utilize the two Evolutionary Algorithms NSGA-II [14]and FEP [15]for the following objectives:

1) NSGA-II was utilized for generating various intelligent computer players.
2) FEP was utilized in order to find an appropriate combination of the values of the three parameters $W_G$, A, and $W_0$.

Thanks to NSGA-II and FEP, we could succeed in finding the following combination of the parameter values: Wg = 0.27, A = 0.04, $W_0$ = 0.12

Due to space, we don't go into details. Interested readers are referred to [13].

## 4 Game Playing Experiments for an Appropriate Evaluation of the Three Games

We have carried out several game playing in order to get an answer concerning to the question "Which game is the best for letting people consider seriously about the commons?". In the followings, we show several data relating to game playing of the three games (which has recently been carried out in our university (Osaka Kyoiku University). Fig.2 illustrates the changes of the total points of the 6 players in each game run. Table 1 shows the total number of the changes of the rankings of the 6 players during each game run.

From Fig.2 and Table 1, we can easily observe:

1) Difference between the total point gained by the top player and that gained by the last player is almost always the smallest in the game playing of the modified COMMONS GAME (constructed by the use of the two Evolutionary Algorithms & NNs) among those of the three games.
2) The number of the times of the changes of the rankings of the 6 players is the highest in the game playing of the modified COMMONS GAME (constructed by the use of the two Evolutionary Algorithms & NNs) among those of the three games.

There are several indicators which can evaluate player's involvement in the game playing. Among those, the two indicators having been utilized above (1)&2)) might

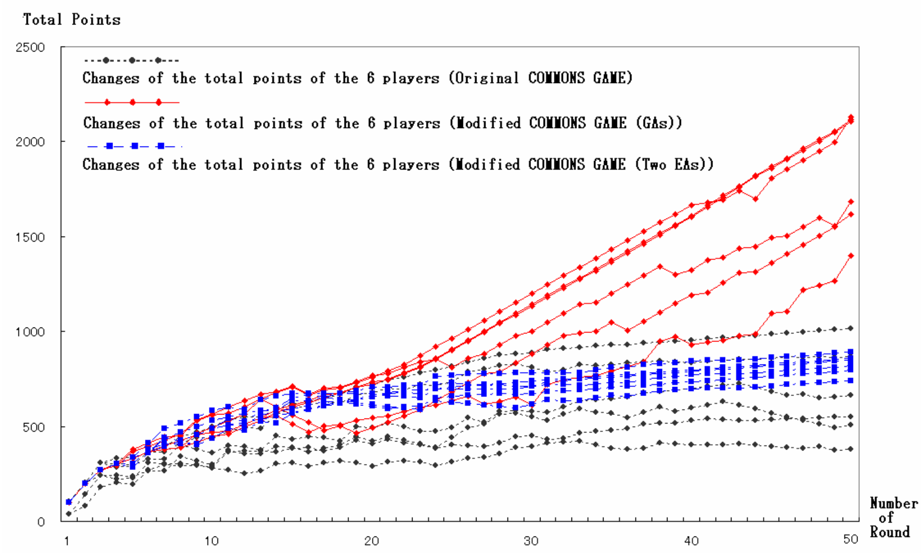


**Fig. 2.** Changes of the total points of the 6 players in each game run

**Table 1.** Total number of the changes of the rankings

| Type of the GAME | Total Number |
|---|---|
| Original COMMONS GAME | 27 |
| Modified COMMONS GAME (GAs & NNs) | 32 |
| Modified COMMONS GAME (two EAs) | 45 |

be by far the two of the most important. After the game playing, one of the authors asked the players concerning their satisfaction about the game playing of the three games. Almost all of the players expressed their positive impression about the game playing of the modified COMMONS GAME (constructed by the use of the two Evolutionary Algorithms & NNs). He also asked the following question: "How do feel about the effectiveness of the game playing from the view point of "letting players consider seriously about the commons" ?" More than half of the players agreed with the effectiveness of the game playing of the modified COMMONS GAME (constructed by the use of the two Evolutionary Algorithms & NNs) toward this purpose.

## 5 Concluding Remarks

In this paper, we have tried to find the best game (from the view point "Which game is the best for letting people consider seriously about the commons ?") among the three games (the original COMMONS GAME, the modified COMMONS GAME utilizing GAs & NNs , and the modified COMMONS GAME utilizing EAs & NNs). Several game playing having been done by our students confirm that the modified COMMONS GAME utilizing EAs & NNs can provide the best chance for letting players consider seriously about the commons. However, this has been confirmed only by several game playing by our students. Future research is needed to carry out lots of game playing by various people for the full confirmation of our research.

Further, continuous effort is also needed for finding a more advanced game for letting people have a chance to consider seriously about the commons.

## Acknowledgements

The authors would like to express their heartfelt thanks to the Foundation for Fusion of Science & Technology (FOST) who has given them financial support.

## References

1. Shubik, M.: Games for Society Business and War: Towards a Theory of Gaming. Elsevier, Amsterdam (1975)
2. Hausrath, A.: Venture Simulation in War, Business, and Politics. McGraw-Hill, New York (1971)
3. Duke, R.: Gaming: The Future's Language. Sage Publications, Thousand Oaks (1974)
4. Powers, P., Duss, R., Norton, R.: THE COMMONS GAME Manual. In: IIASA (1980)
5. Ausubel, J.: The Greenhouse Effect: An Educational Board Game. Instruction Booklet. IIASA (1981)
6. Baba, N., Uchida, H., Sawaragi, Y.: A Gaming Approach to the Acid Rain Problem. Simulation & Games 15(3), 305–314 (1984)
7. Baba, N., et al.: Two Microcomputer-Based Games. In: IIASA Collaborative Paper, WP 86-79, pp. 1–46 (1986)
8. Baba, N.: PC-9801 Personal Computer Gaming System. Nikkan Kogyo Publishing Company (1986) (in Japanese)
9. Baba, N.: An Application of Artificial Neural Network to Gaming. Proceedings of SPIE 2492, 465–476 (1995) (Invited Paper)
10. Baba, N., Kita, T., Takagawara, Y., Erikawa, Y., Oda, K.: Computer Simulation Gaming Systems Utilizing Neural Networks & Genetic Algorithms. Proceedings of SPIE 2760, 495–505 (1996) (Invited Paper)
11. Baba, N.: Application of Neural Networks to Computer Gaming. In: Tzafestas, S.G. (ed.) Soft Computing in Systems and Control Technology, ch. 13, pp. 379–396. World Scientific, Singapore (1999)
12. Baba, N., Jain, L.C. (eds.): Computational Intelligence in Games. Springer, Heidelberg (2001)
13. Baba, N., Jain, L.C., Handa, H.: Advanced Intelligent Paradigms in Computer Games. Springer, Heidelberg (2007)
14. Deb, K., et al.: A Fast and Elitist Multi-Objective Genetic Algorithm: NSGA-II. IEEE Trans. EC 6(2), 182–197 (2002)
15. Yao, X., Liu, Y., Lin, G.: Evolutionary Programming Made Faster. IEEE Trans. EC 3(2), 82–102 (1999)

# Petri Net-Based Simulation and Analysis of the Software Development Process

Gordan Topic[1], Dragan Jevtic[2], and Marijan Kunstic[2]

[1] Ericsson Nikola Tesla d.d, Krapinska 45,
10000 Zagreb, Croatia
gordan.topic@ericsson.com
[2] Faculty of electrical engineering and computing, Unska 3,
10000 Zagreb, Croatia
dragan.jevtic@fer.hr

**Abstract.** This paper explores improvements which can be achieved by applying Petri nets to the modeling, simulation and analysis of the software development process. The huge complexity of this process, in conjunction with the demand for rapid reaction to market pressure, hard limits on time and cost, and the fluidity of human resource organization can make it considerably difficult to establish a confident software development process. Simulations of such processes using Petri net models show considerable benefits with respect to real factors such as resource requirements, representation of critical borders, the effects of resource deficit, delays in process phases, etc.

**Keywords:** software development, process modeling, colored Petri net.

## 1 Introduction

Software development is a complex process, which requires precise planning and realization to meet specific requests. A large portion of software development belongs to the telecommunications sector in which specific software components, which can be broadly dispersed, act together to deliver a requested service.

However, software is like a living being, continuously absorbing good and bad characteristics from its creators, particularly the organizational skills and competences of the development team. Furthermore, it does not tolerate forced development, e.g. urgent delivery. Good and high quality software is a reflection of a competent, organized and satisfied team. Thus, it seems reasonable to invest effort into quality assurance. At times, a manager must react quickly to satisfy changing market demands. However, manufacturing and business processes require an optimal time plan, adjusted to specific conditions. It is often necessary to predict the time required for software development, together with team proportions, while maintaining high quality of the software product. Various methods to discover the process dynamics have been proposed mostly based on modeling and simulation. In this paper, we present an approach to software development process (SWDP) modelling, simulation and analysis

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 418–425, 2008.

using Colored Petri Nets (CPN). Potential benefits lie in process design and analysis, and in organizational improvements prior to the SWDP and during its various phases.

The software development process based on the waterfall model frequently used in telecommunication industry is described in section 2. A model of the SWDP realized using colored Petri nets, along with a short review of Petri nets is given in section 3. Results of simulation experiments and process analysis are elaborated upon in section 4, followed by a conclusion.

## 2 The Software Development Process

Complex information systems as manufactured goods require a high organizational level for software development. In practice, small companies habitually consider direct writing of program code in their software development. However, real software development implies much more and high quality software development requires a high level of business organization and sometimes very complex processes of software production.

### 2.1 Phases of Software Development

The Waterfall_model is document-driven where each step yields artifacts in the form of documents. According to the waterfall model of the SWDP [3], a simplified SWDP can be defined as a business process consisting of four phases. The first phase is *analysis.* The *analysis* phase starts with gathering requirements into an input document for the SWDP. It consists of an exact definition of the customer or client requirements for the requested software. The start of this phase is marked by a decision point, or so-called tollgate, labeled as TG0 (Fig. 1).

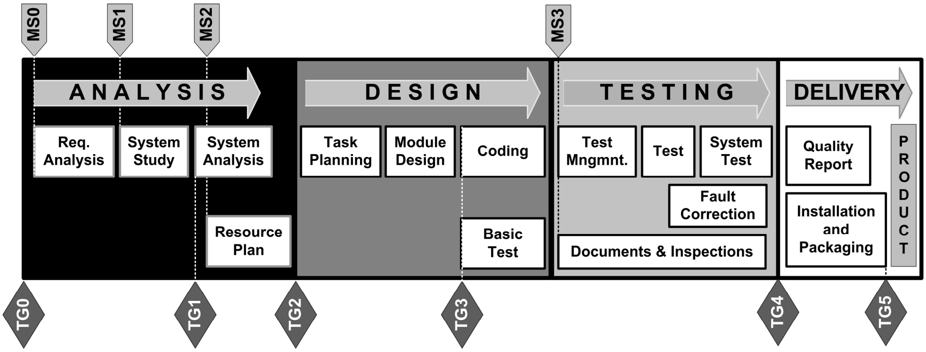


**Fig. 1.** A rough representation of the SWDP depicted in four phases with sub-processes

This phase terminates with tollgate TG2, where a major decision must be made according to the progress of this phase. The decision can be based on the results of requirement analysis, costs estimation and/or a system study, i.e. in sub-processes of this phase. If the progress of the SWDP is satisfactory, the next phase, called the *design* phase, begins. In this phase, the requirements are decomposed and broadly analyzed by performing *task planning*, *module design*, *coding* and *basic testing* (fig. 1). The results are program units, i.e. software modules. The *design* phase terminates by concurrent execution of encoding and basic tests for all software modules. The next control point, the *Milestone* labeled MS3, marks the beginning of the *testing* phase. Software modules have to be tested in a simulated functional environment, along with the related hardware, which is commonly developed as a parallel activity [1]. Tollgate TG4 marks the start of the *delivery* phase of the SWDP, which encompass different final activities, such as product packing, creating an install shield, writing user documentation and quality reporting (Fig. 1). The last part of the SWDP is marked by tollgate TG5. The main phases of the SWDP follow the arrangement shown in fig. 1 in theory. However, many of these process activities are of concurrent nature, i.e. are executed in parallel with various interconnections.

### 2.2 Organization, Structure and Documentation

The SWDP incorporates a well-organized project structure, which must be skillfully managed according to human resource potential. This section describes the organization and layered structure of the SWDP, which we modeled, simulated and analyzed using the CPN model.

Three management levels exist in the SWDP hierarchy, each of which provides a corresponding document. The highest level of the management structure consists of managers who care about the fluent flow of software development. This team is comprised of a *project manager* and a *software quality manager*. The project manager manages the entire project team and determines the job dynamics according to predefined objectives. The software quality manager is responsible for monitoring and predicting the quality of the product, as well as maintaining and improving the development process.

A *system analyst* is a highly competent person with adequate experience, placed in-between the high and low levels of the management hierarchy. His knowledge of hardware and software systems must be broad enough to perform sub-process *system analysis* (fig. 1) and to create hardware and software specifications.

Finally, the lowest level of the management team consists of *technical managers*, which are responsible for the technical characteristics of product realization. These include *configuration* and *testing* performed by *software architects* and *hardware engineers*. Their practical knowledge is joined with some special branches depending on the needs of the SWDP, such as design, testing, hardware engineering, etc. Their activity is mainly focused on the specification of product functionality.

After inspecting software documents during the *analysis* phase, software architects create an implementation proposal document, while hardware engineers perform the same task on the hardware side. Software and hardware implementations appear in the documentation as a specification of its functionalities. The configuration manager prepares a configuration plan, while the test manager writes a test plan prerequisite for system testing and functionality verification. As mentioned earlier, in each sub-process activity, the responsible person creates the specific documentation.

# 3 Software Development Modeling and Simulation Using Petri Nets

Modeling business processes provides clear descriptions and better understanding of complex processing phenomena. Process modeling and simulation have been previously used for various process solution estimations, selection of the most suitable cases and resource optimization. A simulation model of the SWDP is an approximate description of a real process or a system of processes. In general, simulation of a model allows for a quantitative analysis and gives answers to "what-if" questions. The SWDP is composed of a large number of interconnected elements, where human resources carry out the essential functions. The focus of our simulations was the usage of human resources and their relation to process dynamics.

## 3.1 A Short Review of Petri Net Models

A Petri net is a mathematical representation of a discrete distributed system. As a modeling language, it graphically depicts the structure of a distributed system as a directed bipartite graph with annotations. A Petri net consists of places, transitions, and directed arcs. Arcs run between places and transitions. Places may contain any number of tokens. Execution of Petri nets is non-deterministic, i.e., multiple transitions can be enabled at the same time, making Petri nets well-suited for modeling the concurrent behavior of distributed systems [2, 5].

The formal definition of a Petri net describe it as a 5-tuple ($S$, $T$, $F$, $M_0$, $W$), where:

- $S$, $T$ and $F$ are the set of places, transitions, and directed arcs, respectively.
- Set F is subject to the constraint that no arc may connect two places or two transitions, or more formally: $F \subseteq (S \times T) \cup (T \times S)$
- $M_0 : S \rightarrow \mathrm{N}$ is an initial marking, where for each place $s \in$ S, there are $n_s \in$ N tokens.
- $W : F \rightarrow \mathrm{N}^+$ is a set of arc weights, which assigns to each arc $f \in$ F some $n \in \mathrm{N}^+$ denoting how many tokens are consumed from a place by a transition, or alternatively, how many tokens are produced by a transition and put into each place.

Colored Petri nets are a high-level extension of basic Petri nets which support hierarchical structuring, timed transition executions, and where every token has a value. Token interaction must be defined and associated to each transition. Each token can carry a time stamp denoting the time at which the token is ready. A hierarchical process structure can be constructed by substitution transitions of so-called *sockets* and *ports* interconnecting the Petri net on a lower level for that transition. Fusion places are level independent. Once created, a fusion place interconnects related Petri nets on different levels. These properties annotate CPN as an appropriate model for systems in which communication, synchronization and resource sharing are fundamental [4].

### 3.2 A CPN Model of the Development Process

Our CPN model of the SWDP has a hierarchical structure created from a total of 19 colored Petri nets which are arranged to correspond to the SWDP hierarchy in three levels: level 1 is entitled the *process level*, level 2 the *sub-process level* and level 3 the *operative level* (Fig. 2). Each lower level is composed as set of Petri subnets. Subnets on level 2 are driven by and connected to their unique substitution transitions from the upper level. In reality, the execution of a substitution transition on level 1 activates the execution of a subnet on level 2 and then returns to level 1. Analogously, the part of the transition set on level 2 is exchanged by subnets from level 3. Generally, each substitution transition is based on the Petri net of a lower level. The global hierarchical model is illustrated in fig. 2.

In our SWDP model, we used a timed CPN consisting of 120 places, of which 25 fusion and 41 socket places. The model contained 74 transitions, of which 18 substitution transitions were included. The number of simple colors declared was 4: 2 compound and 2 colors with time dimension. The total number of variables that support token movement was 10.

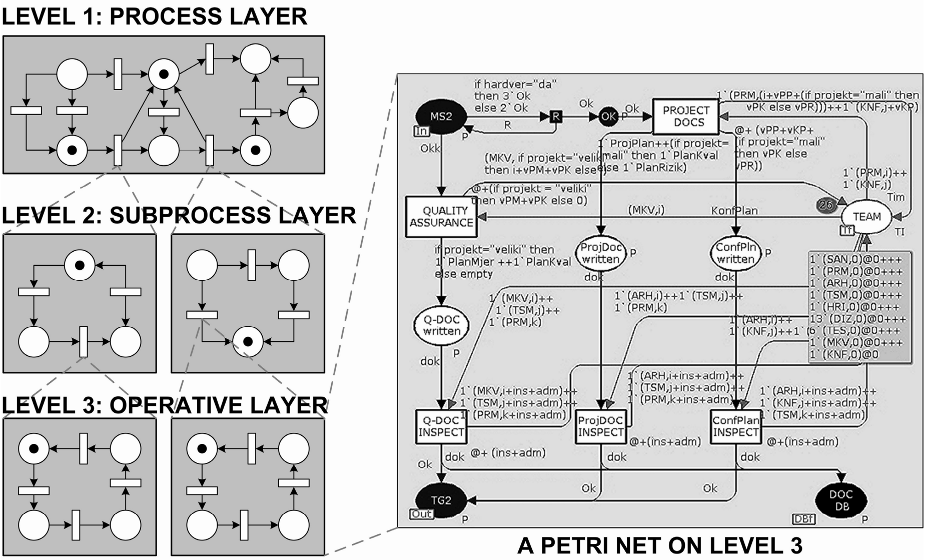


**Fig. 2.** An illustration of the applied CPN model structure for the SWDP arranged in three levels

The simulation model was constructed to provide information regarding human resource allocation, which can be defined as a certain number of team members allocated in 9 different project roles. Changing the initial conditions in the CPN model corresponds to a change in the number of team members in the SWDP. The initial conditions of the simulation model are variables for project specifications, human resources, project documentation, auxiliary and time variables for different process phases. Many simulations were performed by changing the initial conditions of the

CPN model, which manifested in results containing different numbers of members in the project team.

### 3.3 Process Modeling and Data Gathering

Modeling and simulation using the CPN model can, in general, be used for process learning, performance improvement and cost reduction in the SWDP. Our simulations were focused on the optimization of human resources subject to certain financial and time constraints.

The most complicated part of process modeling is gathering the data needed for the simulation. Our data sources were derived and collected from project documentation, quality measuring results, observation and measurements of process activities, and from own experiences as project team members from other parallel or finished projects. The project and time plans, role, and process description documents can provide enough information for fundamental estimation. However, additional information must be collected and extracted from quality measurement records and through empirical approaches.

## 4 Results of Simulation Experiments

As mentioned earlier, documentation in the SWDP was used for transferring knowledge from an idea to a real software product. The SWDP described in this paper, operates with 22 different project documents. The expected benefit of our simulation was to discover the time and cost function for the modeled process regarding human resource usage, which we consider the most important factor in the SWDP. The target was to determine the bordering cases for which project requirements could be satisfactorily managed. The SWDP was defined as a project with the following limitations:

- Project duration for the SWDP could not exceed 2.000 hours, i. e. approximately one year.
- The cost could not exceed a value of 40.000 man/hours, with minimum usage of human resources.

Process modeling and simulation were performed by CPNtools [4]. Initial values of the variables in the CPN represented team members corresponding to the number of designers and testers in the development team. A hundred separate simulations were performed with various initial input variables for designers and testers. The total number of members in the entire development team was defined as the sum of the designers, testers and other members that were recognized as essential to the team. All simulations resulted in a prediction of the amount of time needed for project duration and the effective time interval needed for software product development.

The graphs in fig. 3 show the relation between project duration, the number of designers and testers, and the cost obtained via simulation. The graph on the left shows the solution space in which various numbers of team members could complete the project in the proposed time period. Considering the cost criteria, i.e. a maximum of 40.000 man/hours, significantly cuts the surface area and additionally restricts the

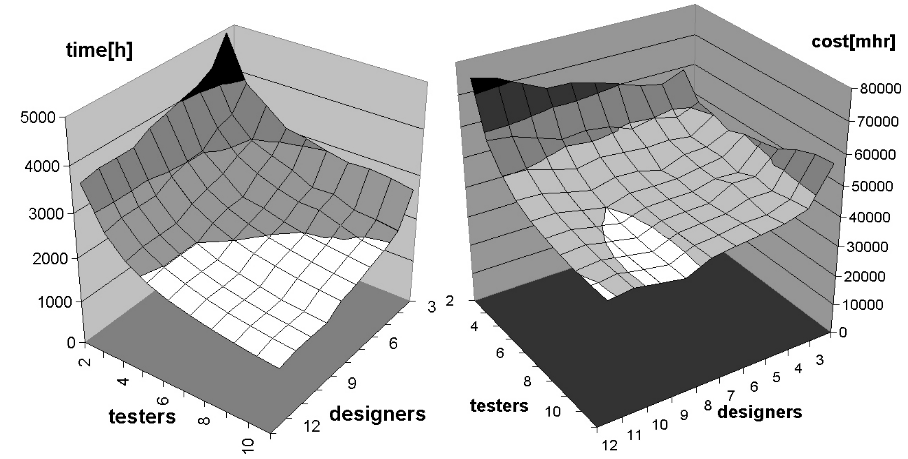


**Fig. 3.** The graph on the left shows project duration, while the graph on the right shows cost estimation, both in relation to the number of designers and testers in the development team

number of project team members (represented by the white areas on both diagrams in Fig. 3). The acceptable parts of the surfaces in the graphs with respect to the defined project conditions are shown in white. The optimal number of members was found for the given structure of the SWDP. Namely, the optimal team consisted of 25 members, i. e. 9 designers, 9 testers and 7 mandatory members, such as quality managers, which appear in big projects.

Such a team could develop the requested software product effectively during 1511 h and with minimal cost of 37 775 man/hours. Based on this result, team member load was calculated for all process roles, depicted in fig. 4. According to the graph, designers consume the largest part of total development time (691,7 hours/man), followed by testers (600,4 hours/man).

The dynamics of time consumption over the SWDP phases is depicted in fig.5. It shows the time distribution from phase to phase, where the increment between phases represents amount of time used (Fig. 5). The most time, in this case, was spent in the testing phase.

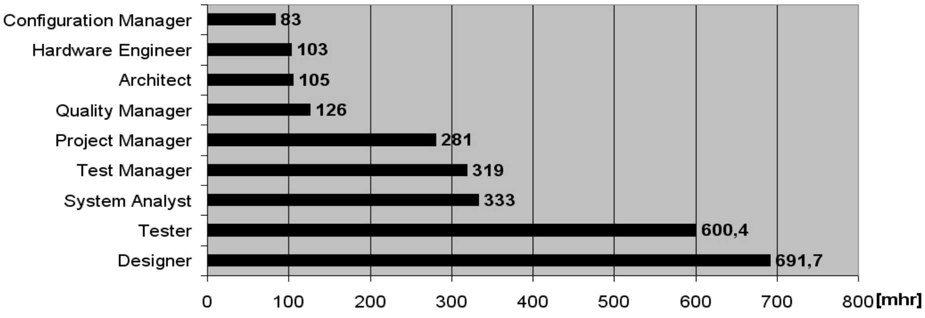
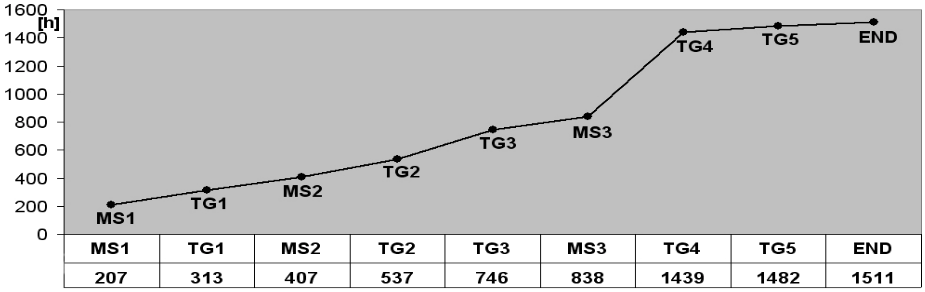


**Fig. 4.** Human resource time spent for different roles in the SWDP



| MS1 | TG1 | MS2 | TG2 | TG3 | MS3 | TG4 | TG5 | END |
|---|---|---|---|---|---|---|---|---|
| 207 | 313 | 407 | 537 | 746 | 838 | 1439 | 1482 | 1511 |

**Fig. 5.** The diagram of time spent over process phases

Additional simulations were performed in which the basic SWDP model was extended to represent tandem projects in order to utilize human resources in periods when human resources were without tasks in the current SWDP. It proved beneficial for total cost saving and increased job effectiveness. The optimal solutions for the SWDP applied to two parallel projects with the same human resources resulted in up to 24% reductions of cost and development time. However, they increased single development time for both projects.

## 5 Conclusion

Any business process can be realized and verified with an equivalent Petri net model. Thus, decision-making and management can be additionally supported with simulation and analysis of such models for SWDP pre-planning and, moreover, during its phases. The main obstacle in process modeling is finding apposite data in advance, which may well represent the process entities for which existing requirements have to meet available resources. This can be, for example, the amount of time needed for an expert to solve the problem. On the other hand, the simulation results provide various allocations of SWDP resources, which decision-makers and management could consider as relevant selections for the observed SWDP.

## References

1. Perry, W.E.: Effective Methods for Software Testing, 3rd edn. John Wiley & Sons, Chichester (2006)
2. Desel, J., Juhás, G.: What Is a Petri Net? Informal Answers for the Informed Reader. In: Ehrig, H., Juhás, G., Padberg, J., Rozenberg, G. (eds.) APN 2001. LNCS, vol. 2128, pp. 1–25. Springer, Heidelberg (2001)
3. Jacobson, I., Booch, G., Rumbaugh, J.: The Unified Software Development Process. Addison-Wesley, Reading (1999)
4. Jensen, K.: Coloured Petri Nets – Basic Concepts, Analysis Methods and Practical Use, vol. 1, 2. Springer, Heidelberg (1996)
5. Peterson, J.L.: Petri net theory and the modeling of systems. Prentice-Hall, Englewood Cliffs (1981)

# Learning Grouping and Anti-predator Behaviors for Multi-agent Systems

Koichiro Morihiro[1,2], Haruhiko Nishimura[3],
Teijiro Isokawa[2,4], and Nobuyuki Matsui[2,4]

[1] Hyogo University of Teacher Education, Hyogo 673-1494, Japan
mori@info.hyogo-u.ac.jp
[2] Himeji Institute of Technology, Hyogo 671-2201, Japan
[3] Graduate School of Applied Informatics, University of Hyogo,
Hyogo 650-0044, Japan
haru@ai.u-hyogo.ac.jp
[4] Graduate School of Engineering, University of Hyogo, Hyogo 671-2201, Japan
isokawa@eng.u-hyogo.ac.jp, matsui@eng.u-hyogo.ac.jp

**Abstract.** Several models have been proposed for describing grouping behavior such as bird flocking, terrestrial animal herding, and fish schooling. In these models, a fixed rule has been imposed on each individual a priori for its interactions in a reductive and rigid manner. We have proposed a new framework for self-organized grouping of agents by reinforcement learning. It is important to introduce a learning scheme for developing collective behavior in artificial autonomous distributed systems. This scheme can be expanded to cases in which predators are present. In this study we integrate grouping and anti-predator behaviors into our proposed scheme. The behavior of agents is demonstrated and evaluated in detail through computer simulations, and their grouping and anti-predator behaviors developed as a result of learning are shown to be diverse and robust by changing some parameters of the scheme.

## 1 Introduction

The collective behavior of creatures can often be observed in nature. Bird flocking, terrestrial animal herding, and fish schooling are the typical well-known cases. Several observations suggest that there are no leaders in such groups who control the behavior of the group. Collective behavior develops from the local interactions among agents in groups [1,2,3]. Several models have been proposed to describe grouping behavior. In these models, a fixed rule has been imposed on each agent a priori for its interactions [4,5,6,7]. This reductive and rigid approach is suitable for modeling groups of biological organisms since they appear to inherit the ability to form groups. However, it is important to introduce a learning scheme that develops collective behavior in artificial autonomous distributed systems.

The characteristic feature of reinforcement learning [8,9] is unsupervised learning by trial and error, i.e., by exploration, in order to maximize rewards obtained

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 426–433, 2008.

from the environment. Introducing appropriate relations between an agent's behavior (action) and its reward, we could design a new scheme for the development of grouping behavior by reinforcement learning. We have proposed a new framework for self-organized grouping of agents by reinforcement learning [10]. This scheme can be expanded to cases in which predators are present [11].

In this study we integrate grouping and anti-predator behaviors into our proposed scheme. The behavior of agents is demonstrated and evaluated in detail through computer simulations, and their grouping and anti-predator behaviors developed as a result of learning are shown to be diverse and robust by changing some parameters of the scheme.

## 2 Reinforcement Learning

Reinforcement learning originated from the experimental studies on learning in the field of psychology. Almost all reinforcement learning algorithms are based on the estimation of value functions. Computer systems receive only an evaluative scalar feedback for a value function from their environment and not an instructive feedback as in supervised learning. Q-learning [12] is known as the best-understood reinforcement learning technique. A value function in Q-learning consists of values determined from a state and an action, which is called Q-value. In Q-learning, the learning process consists of acquiring a state ($s_t$), deciding an action ($a_t$), receiving a reward ($r$) from an environment, and updating the Q-value ($Q(s_t, a_t)$). The Q-value is updated by the following equation:

$$Q(s_{t+1}, a_{t+1}) = Q(s_t, a_t) + \alpha[r + \gamma \max_{a' \in A(s')} Q(s', a') - Q(s_t, a_t)] \quad , \qquad (1)$$

where $A$ denotes the set of actions; $\alpha$, the learning rate ($0 < \alpha \leq 1$); and $\gamma$, the discount rate ($0 < \gamma \leq 1$). Q-learning is one of the reinforcement learning techniques used for maximizing the sum of the rewards received. It attempts to learn the optimal policy by compiling a table of Q-values $Q(s, a)$ according to Eq. (1). $Q(s, a)$ provides the estimated value of the expected response action $a$ for state $s$. Once these Q-values are learned, the optimal action for a state is the action with the highest Q-value. In the original Q-learning algorithm, a greedy policy with pure exploitation has been adopted. However, it is generally difficult to obtain satisfactory results by employing this policy. Therefore, in the present study, a policy that allows the adoption of a nonoptimal action is introduced.

In reinforcement learning, many types of exploration policies have been proposed for learning by trial and error, such as $\epsilon$-greedy, softmax, and weighted roulette action selection. In the present study, we adopt the softmax action selection method, and the rule is given as $p(a|s) = \frac{\exp\{Q(s,a)/T\}}{\sum_{a_i \in A} \exp\{Q(s,a_i)/T\}}$ , where $T$ is a positive parameter called temperature. High temperatures cause all the actions to be (nearly) equiprobable.

# 3 Model and Method

## 3.1 Internal Perceptual Space of Each Agent

We employ a configuration of $N$ agents that can move in any direction in a two-dimensional field. Learning of each agent (agent $i$) progresses asynchronously with time in the discrete time step $t_i = d_i t + o_i$, where $d_i$ and $o_i$ are integers proper to agent $i$ $(0 \leq o_i < d_i)$. The agents act in discrete time $t$, and at each time step $t_i$ an agent (agent $i$) finds another agent (agent $j$) among $N-1$ agents and learns.

In the internal perceptual space, state $s_t$ of $Q(s_t, a_t)$ for agent $i$ is defined as $[R]$, which is the maximum integer not exceeding the Euclidean distance $R$ from agent $i$ to agent $j$. For action $a_t$ of $Q(s_t, a_t)$, four types of action patterns ($a_1$,$a_2$,$a_3$, and $a_4$) are considered, which are as follows (also illustrated in Fig. 1):

$a_1$ : Attraction to agent $j$
$a_2$ : Parallel positive orientation to agent $j$ $(\mathbf{m_a} \cdot (\mathbf{m_i} + \mathbf{m_j}) \geq 0)$
$a_3$ : Parallel negative orientation to agent $j$ $(\mathbf{m_a} \cdot (\mathbf{m_i} + \mathbf{m_j}) < 0)$
$a_4$ : Repulsion to agent $j$

Here, $\mathbf{m_a}$ is the directional vector of $a_t$, and $\mathbf{m_i}$ and $\mathbf{m_j}$ are the velocity vectors of agents $i$ and $j$, respectively. Agent $i$ moves in accordance with $\mathbf{m_i}$ in each time step, and $\mathbf{m_i}$ is updated by the expression

$$\mathbf{m_i} \leftarrow (1-\kappa)\mathbf{m_i} + \kappa\mathbf{m_a} \quad , \tag{2}$$

where $\kappa$ is a positive parameter $(0 \leq \kappa \leq 1)$ called the inertia parameter.

In this study, as we consider same types of agents and the perceived object as a predator, two types of corresponding Q-values should be introduced.

## 3.2 Learning Modes against Agents of the Same Type and against Predators

In our proposed model, we offer the reward $r$ for $Q(s_t, a_t)$ to each agent according to the distance $R$ from the perceived agent of the same type. The learning of

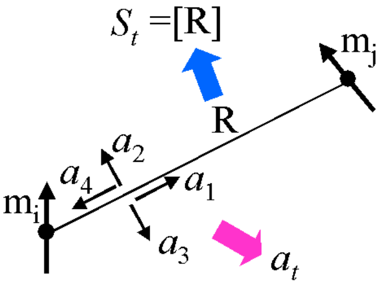


**Fig. 1.** Constitution of internal perceptual space of each agent

**Table 1.** Reward $r$ for selected action $a_t$ in state $s_t = [R]$

| | Learning mode against agents of the same type | | | | | | | Learning mode against predators | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $s_t$ | $0 < [R] \leq R_1$ | | $R_1 < [R] \leq R_2$ | | $R_2 < [R] \leq R_3$ | | $R_3 < [R]$ | $0 < [R] \leq R_3$ | | $R_3 < [R]$ |
| $a_t$ | $a_4$ | $a_1, a_2, a_3$ | $a_2$ | $a_1, a_3, a_4$ | $a_1$ | $a_2, a_3, a_4$ | $a_1, a_2, a_3, a_4$ | $a_4$ | $a_1, a_2, a_3$ | $a_1, a_2, a_3, a_4$ |
| $r$ | 1 | $-1$ | 1 | $-1$ | 1 | $-1$ | 0 | 1 | $-1$ | 0 |

the agents proceeds according to the positive or negative reward, as shown in Table 1, in which $R_1 < R_2 < R_3$. In the case of $0 < [R] \leq R_3$, agent $i$ can perceive another agent of the same type with a probability in proportion to $R^{-\beta}$, where $\beta$ is a positive parameter. This implies that the smaller the value of $R$ is, the easier the selection of the agent at the position is. When $0 < [R] \leq R_1$, the agent receives a positive reward (+1) if it assumes a repulsive action against the perceived agent ($a_4$); otherwise, it receives the penalty ($-1$). In the case of $R_1 < [R] \leq R_2$ and $R_2 < [R] \leq R_3$, the agent also receives the reward or penalty defined in Table 1 depending on the actions. In the case of $[R] > R_3$, agent $i$ cannot perceive agent $j$; hence, it receives no reward and chooses an action from the four action patterns ($a_1$, $a_2$, $a_3$, and $a_4$) randomly.

When there is a predator within $R_3$, agent $i$ perceives the predator with a probability of 1, and the learning mode against agents of the same type is switched to the learning mode against predator. In this case, agent $i$ gets the positive reward (+1) if it takes a repulsive action to evade the predator ($a_4$); otherwise, it gets the penalty ($-1$), as defined in Table 1.

## 4 Simulations and Results

In the computer simulations, we have assumed the following experimental conditions: $\alpha = 0.1$, $\gamma = 0.7$ in Eq.(1), $T = 0.5$ (under learning) for the softmax action selection method, $\kappa = 0.4$ in Eq.(2), $\beta = 0.5$ for the distance dependence of $R^{-\beta}$, $d_i = 1$, and $o_i = 0$. The initial velocities of the same type of agents are set to one body length (1 BL). The velocity $|\mathbf{m_a}|$ which is the directional vector of $a_t$ is also set to one body length (1 BL). The velocity of the predator is set to two body lengths (2 BL). We have simulated our model for the number of agents $N = 30$ and $R_1 = 4$ (BL), $R_2 = 20$ (BL), and $R_3 = 50$ (BL).

### 4.1 Evaluation in No Predator Case and in the Case Predator Appears

In order to quantitatively evaluate how the agents develop grouping behavior, we introduce the measure $|\mathbf{M}|$ of the uniformity in direction and the measure $E$ of the spread of agents.

$$|\mathbf{M}| = \frac{1}{N} \left| \sum_{i=1}^{N} \mathbf{m_i} \right| , \tag{3}$$

$$E = \frac{1}{N}\sum_{i=1}^{N}\sqrt{(x_{\mathrm{A}}^{i} - x_{\mathrm{G}})^2 + (y_{\mathrm{A}}^{i} - y_{\mathrm{G}})^2} \tag{4}$$

where $(x_{\mathrm{A}}^{i}, y_{\mathrm{A}}^{i})$ and $(x_{\mathrm{G}}, y_{\mathrm{G}})$ are the two-dimensional coordinate of agent $i$ and the barycentric coordinate among the agents, respectively. The value of $|\mathbf{M}|$ becomes closer to 1 when the directions of agents increase their correspondence. The agents come close when the value of $E$ becomes small. In the evaluation, we take 100 events of simulation with various random series in exploration in both no predator case and the case predator appears.

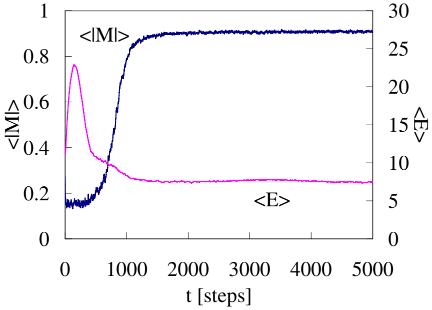


**Fig. 2.** Time step dependence of averaged $|\mathbf{M}|$ and $E$ in 100 events for no predator case

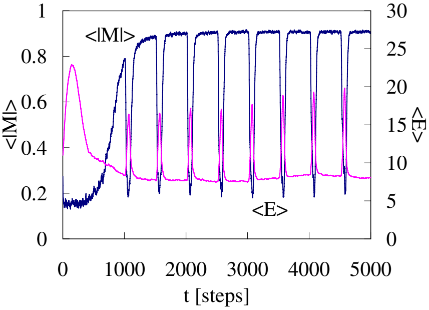


**Fig. 3.** Time step dependence of averaged $|\mathbf{M}|$ and $E$ in non-splitting 94 events for the case predator appears

Figure 2 shows the time step dependences of averaged $|\mathbf{M}|$ and $E$ for no predator case. The transition of $\langle|\mathbf{M}|\rangle$ evolves good in every time step. The value of $\langle E\rangle$ takes a large value at the early stage of learning, after which it decreases to a value around 8 as learning proceeds.

In the case predator appears, the predator approaches the agents from behind and passes straight through the center of the group of agents. The predator appears in every 500th time step up to 5000 time steps. Figure 3 shows the average of non-splitting 94 events in 100 events. When the predator appears, the learning mode is changed. Hence, $\langle E\rangle$ takes a large value and $\langle|\mathbf{M}|\rangle$ decreases to around 0.2. This implies that the agents do not exhibit grouping behavior. When the predator disappears, the learning mode is reverted to the original mode. $\langle E\rangle$ takes a small value and $\langle|\mathbf{M}|\rangle$ increases again to around 0.9 because of the grouping behavior exhibited by the agents.

### 4.2 Effect of Inertia Parameter on Grouping in No Predator Case

From the definition of updating the velocity vector of an agent $\mathbf{m_i}$ (Eq.(2)), an agent has stronger inertia (tendency to keep its own direction unchanged) when $(1-\kappa)$ takes a larger value. Figure 4 shows $(1-\kappa)$ dependences of $\langle|\mathbf{M}|\rangle$ and $\langle E\rangle$ at the end of learning ($t = 5000$). The spread of agents $\langle E\rangle$ becomes considerably large and the directional uniformity of agents $\langle|\mathbf{M}|\rangle$ becomes lower when $(1-\kappa)$ exceeds 0.6. This means that there grow two (or more) groups of agents, due to breakup of a group. Because agents with strong inertia (large $(1-\kappa)$) need some time steps to change their directions according to other agents, they sometimes cannot keep track of other agents, and several agents get segregated. Figure 5 shows the case under the condition that the velocities $|\mathbf{m_i}|$ of all agents are fixed to 1 ($\mathbf{m_i} \leftarrow \frac{(1-\kappa)\mathbf{m_i}+\kappa\mathbf{m_a}}{|(1-\kappa)\mathbf{m_i}+\kappa\mathbf{m_a}|}$). Due to the restriction, the threshold $(1-\kappa)$ for breakup of a group becomes 0.5 lower than 0.6 in Fig. 4.

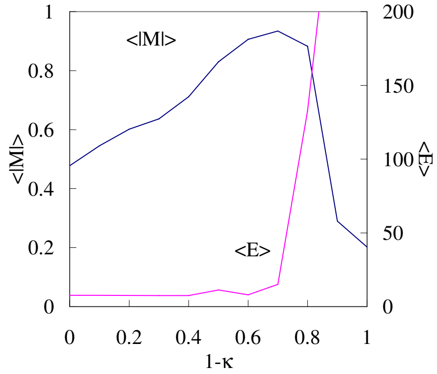


**Fig. 4.** $(1-\kappa)$ dependences of averaged $|\mathbf{M}|$ and $E$ at $t = 5000$ in Eq. (2)

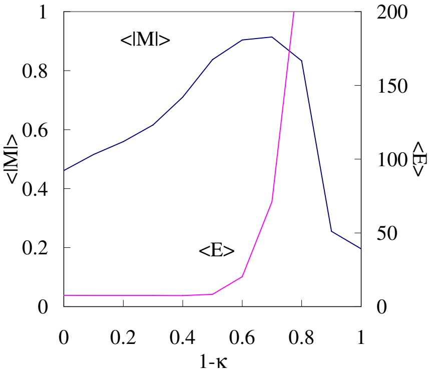


**Fig. 5.** $(1-\kappa)$ dependences of averaged $|\mathbf{M}|$ and $E$ at $t = 5000$ in $|\mathbf{m_i}| = 1$

### 4.3 Velocity Distribution of the Group

The agent changes the velocity $\mathbf{m_i}$ in each time step based on the definition for the velocity vector of an agent. Learning grouping and anti-predator behaviors, the agent also improves its speed for such behaviors. In order to confirm the influence of learning on the velocity of the agent, we check the velocity $|\mathbf{m_i}|$ of the agent under learning and after learning. Figure 6 shows the velocity $|\mathbf{m_i}|$ distribution of the group in no predator case for 500 steps intervals ($500 steps \times 30 agents$ data). In the distribution for 0-500 steps under learning, low speed holds more than 10% and high speed does not reach 40%. In the distribution for 500-1000 steps under learning, low speed decreases to 7% and high speed increases to 52%. At the stage of 1000-1500 steps under learning, the distribution becomes same as that for 500 steps after learning. The velocity $|\mathbf{m_i}|$ distributions of the group when a predator appears are shown in Fig. 7. A similar tendency to Fig. 6 is obtained for 100 steps intervals against the predator.

### 4.4 Trajectories of Agents and Predator

Figure 8 shows the trajectories of the agents in 1700 steps against the predator after learning. In this case, each agent uses fixed Q-value at t=5000 under learning

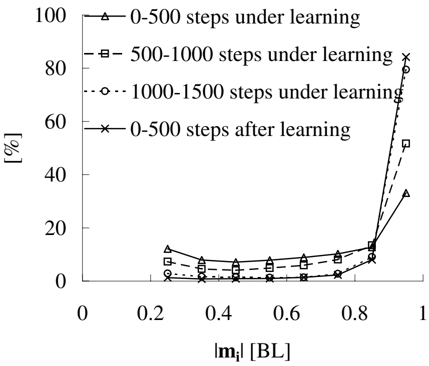


**Fig. 6.** The velocity distribution of the group in no predator case

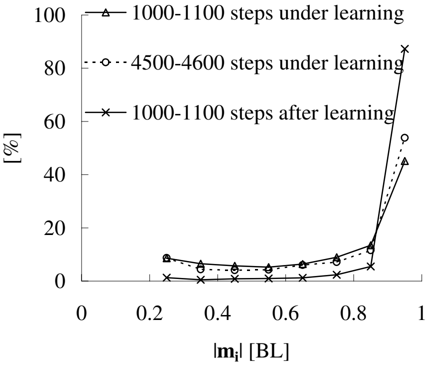


**Fig. 7.** The velocity distribution of the group in the case predator appears

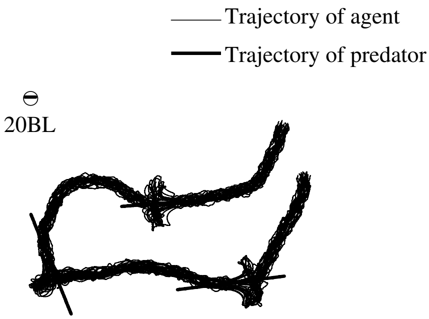


**Fig. 8.** Trajectories of agents in 1700 steps after learning

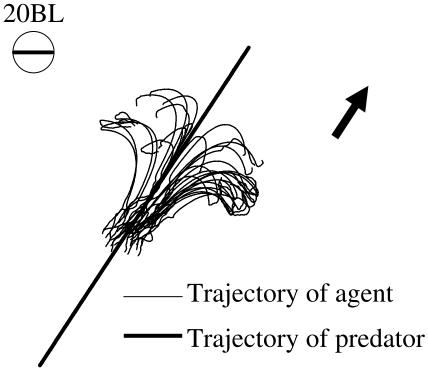


**Fig. 9.** Magnification of Fig. 8 near appearance of predator

and by setting $T \to 0$ in the softmax action selection method as the greedy behavioral policy. Through the learning stages, they have learned grouping and anti-predator behaviors. The magnification of 100 steps in Fig. 8 is shown in Fig. 9. On spotting the predator, the agents form a shape resembling a (polarized) fountain to escape from it. This suggests that the adaptive behaviors of agents, including escaping from the predator, is developed as a result of the two learning modes. Many kinds of anti-predator strategy are observed and recorded from a field study on predator-prey interactions [3,7]. In our simulation, such anti-predator behaviors of agents like vacuole and herd are also observed.

## 5 Conclusion

We have demonstrated a scheme for forming autonomous groups of agents by reinforcement Q-learning. In addition to the grouping behavior of agents, the anti-predator behavior exhibited while escaping from predators can be developed by learning. This indicates the adaptive flexibility of our proposed scheme. In order to confirm effectiveness of our scheme for various situations and patterns of escaping behavior, we have carried out further investigations. We are interested in the examination of the group that has complex and diverse learning conditions. We are carrying out a simulation on a group of agents in which learning progresses asynchronously and on a group that includes slow-learning agents.

## References

1. Shaw, E.: Schooling Fishes. American Scientist 66, 166–175 (1978)
2. Partridge, B.L.: The structure and function of fish schools. Scientific American 246, 90–99 (1982)
3. Pitcher, T.J., Wyche, C.J.: Predator avoidance behaviour of sand-eel schools: why schools seldom split. In: Noakes, D.L.G., Lindquist, B.G., Helfman, G.S., Ward, J.A. (eds.) Predators and Prey in Fishes, pp. 193–204. The Hague, Junk (1983)
4. Aoki, I.: A Simulation Study on the Schooling Mechanism in Fish. Bulletin of the Japanese Society of Scientific Fisheries 48(8), 1081–1088 (1982)
5. Reynolds, C.W.: Flocks, Herds, and Schools: A Distributed Behavioral Model. Computer Graphics 21(4), 25–34 (1987)
6. Huth, A., Wissel, C.: The Simulation of the Movement of Fish Schools. Journal of Theoretical Biology 156, 365–385 (1992)
7. Vabo, R., Nottestad, L.: An individual based model of fish school reactions: predicting antipredator behaviour as observed in nature. Fisheries Oceanography 6, 155–171 (1997)
8. Kaelbling, L.P., Littman, M.L., Moore, A.W.: Reinforcement Learning: A Survey. Journal of Artificial Intelligence Research 4, 237–285 (1996)
9. Sutton, R.S., Barto, A.G.: Reinforcement Learning. MIT Press, Cambridge (1982)
10. Morihiro, K., Isokawa, T., Nishimura, H., Tomimasu, M., Kamiura, N., Matsui, N.: Reinforcement Learning Scheme for Flocking Behavior Emergence. Journal of Advanced Computational Intelligence and Intelligent Informatics(JACIII) 11(2), 155–161 (2007)
11. Morihiro, K., Nishimura, H., Isokawa, T., Matsui, N.: Reinforcement Learning Scheme for Grouping and Anti-predator Behavior. In: Apolloni, B., Howlett, R.J., Jain, L. (eds.) KES 2007, Part III. LNCS (LNAI), vol. 4694, pp. 115–122. Springer, Heidelberg (2007)
12. Watkins, C.J.C.H., Dayan, P.: Q-learning. Machine Learning 8, 279–292 (1992)

# Network Simulations for Relationality Design − An Approach Toward Complex Systems

Katsunori Shimohara

Department of Information Systems Design, Doshisha University
1-3 Tataramiyako-dani, Kyotanabe, Kyoto 610-0321, Japan
kshimoha@mail.doshisha.ac.jp

**Abstract.** We have been conducting research on how to design relationality in complex systems composed of intelligent tangible or intangible, artificial artifacts, by using evolutionary algorithms and network science as methodologies. In addition, we have developed a special machine called Network Simulator as a research tool to conduct massively parallel and ultra-high speed simulations on relationality design. This paper describes the research concept of relationality design and network simulations characterized by automatic hypothesis-finding and verification.

**Keywords:** relationality design, evolutionary algorithms, network analysis, network simulations with hypothesis-finding.

## 1 Introduction

In network science, a system is modeled as a network in which elements of the system are represented by nodes and interactions between elements by edges. The idea to envisage a system as a network can be applied to complex systems at various levels of hierarchy, from molecules, genes and cells, to human organization and society, and economical and social systems.

As a matter of face, recent studies on network analysis of complex systems have revealed the common characteristics for them. For example, properties represented by small world network and/or scale free network have given us a new view to grasp and understand such complex systems as network dynamics. In other words, those systems are supposed to share some common mechanisms to gather, edit and represent information, and to achieve some dynamical functions.

Our research at the laboratory of Socio-informatics, Doshisha University, is oriented towards the design and analysis of complex systems and societies composed of intelligent tangible or intangible, artificial artifacts. The research is based on emphasizing and investigating the role and values of socio-economics' aspects of these systems and societies such as information exchange, interaction, cooperation, psychology and emotions. We are particularly interested in the dynamic properties of adaptive systems and societies such as emergence, growth, development, fusion, fragmentation, and collapse of interaction networks between the artifacts composing these systems.

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 434–441, 2008.

So far we have been conducting research on how to design *relationality* in complex systems by using evolutionary algorithms and network science as methodologies. The concept of *relationality* here denotes *interactions* through which two entities mutually influence each other, *linkage* over time and space, and *context* as a result of accumulated *interactions* and *linkage*. We have also developed a special machine called Network Simulator as a research tool to conduct massively parallel and ultra-high speed simulations on relationality design.

In this paper, we introduce the research concept of relationality design and network simulations characterized by automatic hypothesis-finding and verification.

## 2 Research on "Designing Relationality"

The concept of *relationality* here, again, denotes *interactions* through which two entities mutually influence each other, *linkage* over time and space, and *context* as a result of accumulated *interactions* and *linkage*.

We human beings behave and communicate with others, sometimes based on the past memories and sometimes on anticipation for the future. That is, current humans' information processing is influenced by the past and the future. Thus, it is crucial to consider *linkage* over time and space as well as direct interactions especially in such systems where humans as entities or elements are involved. In that sense, *context* as a result of accumulated interactions and linkage is also very important.

Interactions and linkage form context with the passing of time, and in turn the context and linkage affect upcoming interactions. Thus interactions, linkage and context are mutually related. Relationality is not limited to physical and spatial one, or rather basically invisible and information-driven, and sometimes ecological and environmental. Social and economical systems, culture, region, and sense of value, therefore, are included in relationality.

We human beings should be entities that wish for relationality or relationships with others and hope to find meaning in these relationships. Human beings, in other words, might live in relationality and be alive with relationality. Along with the progress towards informational and networked societies, new social systems in which environmental and information-driven relationality plays an important role will emerge.

The idea to envisage systems, therefore, as relationality networks could be applied to complex systems. The topics of research include analysis of relationality between entities and the resulting emergent properties of complex systems and societies at various levels of hierarchy, from the lowest, molecular level (interactions between molecules in the cells)[1] through the DNA (genetic regulatory networks), cells (interactions between cells during the growth, differentiation and specializations of tissues and organs in multi-cellular organisms) to the highest level - artificial (interactions and collaboration between agents in multi-agent systems) and human societies (human communications and interactions).

Fig.1 shows some of research topics, and examples of research directions include:

- Genome bioinformatics,
- Artificial evolution for developing intelligent software systems [2][3],
- Hardware implementation of developing, evolving and adapting neural networks,
- Artificial emotions,

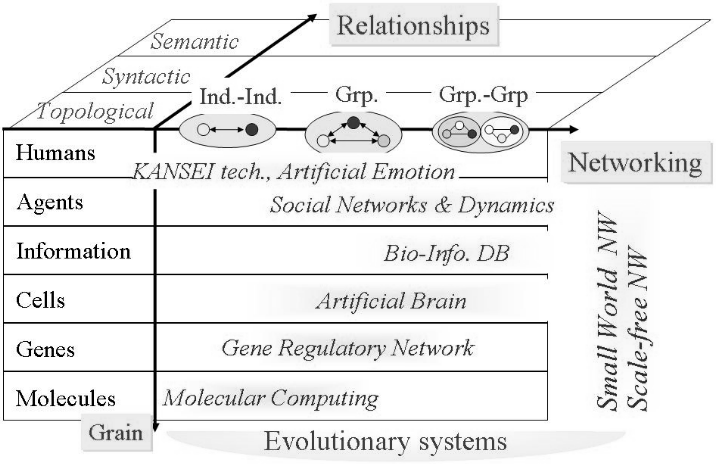


**Fig. 1.** Research topics on relationality design

- Multi-agent systems for modeling and analyzing organizations, societies, economies and their emergent properties.
- Socio-economical aspects of roles and values of entities in human network dynamics.

The approach for design and analysis of such complex systems and societies is based on the algorithmic paradigm of evolutionary algorithms, and especially, genetic programming [4]. In genetic programming, the candidate solutions (represented as genetic programs) to the design problem undergo alterations through genetic operations (such as selection and reproduction) and their survival depends on the fitness (i.e. quality of achieved solution to the problem) tested in the environment, which allows the population of solutions to evolve automatically in a way much similar to the evolution of species in the nature.

As a holistic algorithmic paradigm, evolutionary algorithms are consonant with the holistic approach to the design and analysis of complex systems and societies, based on the belief that any complex system or society is more than the sum of its individual entities, more than the sum of the parts that compose it. And due to their heuristic nature, evolutionary algorithms offer the opportunity to explore various problems in the considered problem domains where the lack of exact analytical solutions or the extreme computational expensiveness of such solutions hinders the efficiency of traditionally applied analytical approaches.

Applying evolutionary algorithms to the design of complex systems, societies, organizations, economics and analysis and comprehension of the emergent properties, and especially the dynamics of interactions between the intelligent tangible or intangible, artificial artifacts composing these systems is the primary objective of our research.

## 3 Network Simulators for Network Simulations

### 3.1 Concept and Key Technologies for Network Simulator

As we take the constructive approach through simulation, a powerful tool for simulations, e.g. a simulator is indispensable in addition to methodologies. We have developed a special machine called Network Simulator for that purpose.

In designing a simulator, one of the concepts is to pursue the essence of complex systems in terms of information processing. That is, it is crucial to emulate relationality of complex systems with high-fidelity as much as possible. So, we could model a system as a complex which accumulates a huge number of relatively simple relationality in the system. As for the architecture of the simulator, therefore, we have employed hardware-oriented architecture, completely different from the direction of supercomputers which make too much of processors' speed.

Another concept we propose here is automatic hypothesis-finding. Usually simulations should be conducted to verify some working hypothesis. In other words, hypotheses should be ready prior to simulations in general. In this research, however, we have proposed simulations with automatic hypothesis-finding. That is, we intend to conduct simulations generating hypotheses and verifying them automatically.

We have developed the Network Simulator that enables not only massively parallel and ultra-high speed simulations for a large-scale system but also simulations with automatic hypothesis-finding. More concretely, this simulator can generate networks as relationality between elements of a given system dynamically on hardware, and then can execute simulations repeating generation and verification of its possible networks automatically.

Key technologies are "flexible hardware", "evolutionary algorithms" and "network analysis". By using reconfigurable hardware, e.g., FPGA (Field Programmable Gate Array), "flexible hardware" becomes available. And then, we can implement, more precisely we can make the simulator implement networks or interactions between elements on the reconfigurable hardware in parallel and dynamically. Secondly, we can change, remake and/or evolve the networks or interactions by using "evolutionary algorithms". The networks or interactions implemented are verified through simulations. That is, this automatic generation and test process of networks or interactions between elements corresponds to the process of hypothesis-finding. In this process, "network analysis" works to effectively reduce the search space for possible interactions so as to maintain network properties as a whole of the system. Thus, good hypotheses should be selected and be genetically modified for the next generation in evolutionary algorithms, while bad ones should be deleted.

### 3.2 Architecture

A new architecture has been developed for an ultra high-speed simulator that computes the evolution of biochemical signal transductions in cells as a typical example of complex systems [5]. Unlike ordinary computers used for most scientific simulations, this architecture is not based on an ALU/FPU type sequential arithmetic process; instead, it directly implements the target phenomena into electronic circuits that operate massively in parallel. This methodology squeezes the potential computing power of the circuits on the silicon far more efficiently than CPU-oriented circuitry. As a result, the cost performance of the developed simulation system is much higher than software simulations operating on the conventional computers. Actually, its performance is comparable to supercomputer systems, while such costs as electricity and placement area are only several times larger than PCs.

From a microscopic viewpoint, the phenomena of life are the results of biochemical reactions. Inside a living cell, about 10 billion biochemical molecules react at a speed of 10 thousand times per second. Molecules are classified into about 20 thousand substances. A substance may be both substrates of reactions or products of other reactions. Accordingly, they form a huge network with the reactions. Knowing the amount of the evolution of substances, one can clarify the phenomena of growth and disease, which in turn are essential for personalized medical care or efficient food production, etc. However, the network is too complex to solve its behavior analytically; only simulation methodology is applicable.

To do so, it is necessary to update the amount of substances 200 million times per second so that the simulation speed matches the real reactions. For more practical case of actual use, for example, one thousand times faster than the actual speed, the number increases to 200 billion per second: a figure 200 times larger than one billion operations per second, which is the typical speed of the floating point arithmetic of ordinary computers. In addition, in software programs, several dozens to several thousand machine clocks are usually needed to update one substance's amount. Therefore, the required computing power is four to six orders larger than PC's.

The main part of the architecture consists of a large number of processing units (PUs); a counter exists in that represents the amount of a substance (Fig.2). It is decreased if representing a substrate of a reaction and decreased if representing a product. These calculations are operated in parallel, and the value of each counter evolves as the simulation progresses. Each reaction may occur randomly, but the average frequency depends on enzyme density and reaction characteristics.

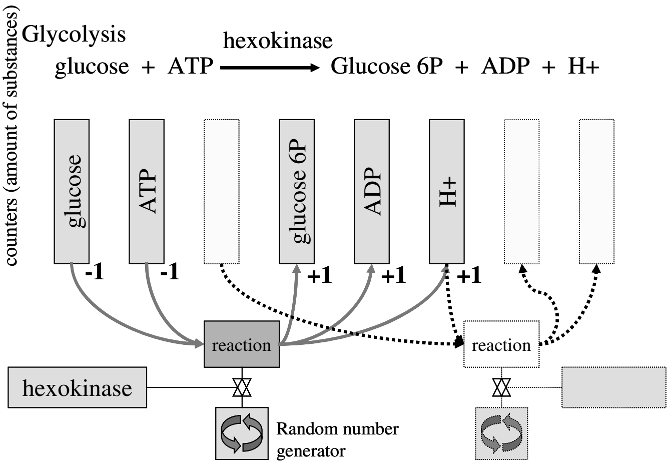


**Fig. 2.** Operations of counters according to the Glycolysis reaction

In the system, these reaction speed controls are achieved by combinations of random number generators and regulations of the degree of incrementing or decrementing. As mentioned above, some counters have to be incremented or decremented repeatedly if they are related to some reactions. In the system, these plural operations are computed simultaneously only in one time step. While providing such useful high level operations, the circuit scale for a PU is still very small due to the simple architecture. The prototype system implements 4096 PU's and operates in 100 MHz. The estimated total performance is up to one million times faster than PC's, sufficiently satisfying the required conditions.

### 3.3 Visualization of Large Scale Gene Interaction Network

We have conducted simulations to simulate and predict gene interaction networks in biological cells [6][7]. Computer simulations of biological phenomena are valuable for system level understanding of biological systems described at molecular level. Drug screening is another application. Simulations of the whole network are important because interactions of individual components are ultimately responsible for an organism's form and functions. Several systems have been proposed for simulation, for instance E-Cell [8], but they lack the execution speed required to simulate real biological organisms, which has a large number of elements and complex interactions among elements.

For instance, a human cell has about 30,000 types of genes and these genes produce substances that participate in a signal transduction network composed of approximately hundred thousand types of elements. Conventional simulation systems are software based, and even running on cluster or grid systems consisting of one thousand processors, completion of simulations in a practical time is unfeasible.

Network Simulator has a novel architecture that is completely different from ordinary computers, designed specifically to simulate models described as large networks. The simulation model consists of quantities associated with elements, and interactions among elements that govern modifications of elements' quantities. The simulator has several thousands of PUs, and is scalable. Each PU stores the quantity of assigned elements, and interactions among elements are simulated by modifying the values that correspond to elements' (substances') quantities.

Preliminary simulations of biochemical reactions indicate clear advantage of Network Simulator over conventional systems, as the simulator is one million times faster than the software simulation for the same simulation precision. Improvement of execution speed by orders of magnitude considerably reduces the turn-around-time of simulation experiments, and opens possibilities for new research methodologies, as numerous parameter sets can be tested simultaneously.

An essential component for simulations of large scale models is a system to visualize comprehensively and succinctly the large amount of generated data, for easier operation of simulation. Furthermore, in the case of gene interaction network, visualization of the gene interaction network under simulation is also necessary.

The high simulation speed of Network Simulator implies that gene interaction networks simulated with Network Simulator have number of nodes that is orders of magnitudes larger than conventional simulators, and conventional visualization mechanisms are inapplicable.

Input data for simulations on Network Simulator is one of high throughput experimental data such as DNA microarray, mRNA and yeast two hybrid experiments. Microarray data is the primary data, and dozens of microarray data are used for a single simulation. Both time sequence data and mutation analysis data or mixed are possible. To enable direct comparison with experimental data, input (experimental) data and simulated data can be shown in microarray-like visualization. Color mapping identical to the experimental result is employed, where the brighter red indicates more abundant quantity, brighter green indicates less quantity, and dark red and green intersect in intermediate amount.

Network Simulator also provides a comprehensive visualization of gene interaction networks under simulation, and its pen display GUI allows intuitive and direct manipulation of simulated networks and simulation parameters (Fig.3). To run a simulation, the user specifies substances, interactions among substances, and initial quantities of substances. Microarray experiment data can also be used as initial values, allowing direct input of microarray experiment results. Gene interaction network is visualized as a network whose topology is optimized according to spring model [9]. During simulations, quantities of substances are displayed in real time as (1) time course quantity curve, and (2) microarray data. While the time course curve is useful to grasp quantity variations, the latter visualization method is useful for direct comparison with microarray experiment results. It is also possible to show multiple microarray displays in sequence. Moreover, microarray display is a collection of squares representing the substances under simulation, and by clicking a square in the microarray window, the system shows detailed information of the substance such as the substance name, biochemical properties, and participating gene interactions.

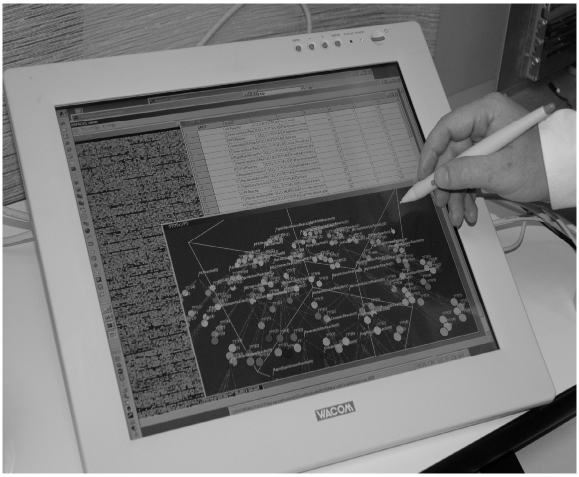


**Fig. 3.** Pen display GUI showing microarray like quantity data (left) and gene interaction network (lower right)

## 4 Conclusion

In this paper, we have proposed the concept of relationality that denotes interactions, linkage over time and space, and context as a result of accumulated interactions and linkage. Also we discussed research direction and issues on how to design relationality in complex systems composed of intelligent tangible or intangible, artificial artifacts, by using evolutionary algorithms and network science as methodologies.

As a research tool, moreover, we developed Network Simulator to conduct massively parallel and ultra-high speed simulations with automatic hypothesis-finding and verification. We applied the simulator to simulation and visualization of large scale gene interaction network, and verified that it works well and is quite effective in computational speed and complexity.

**Acknowledgement.** The author would like to thank Prof. Ivan Tanev, Doshisha University, Japan, Prof. Tetsuya Maeshiro, Tsukuba University, Japan, and Dr. Hitoshi Hemmi, NTT Communication Science Laboratories, Japan for their contribution and cooperation to this research as colleagues.

## References

[1] Liu, J.Q., Shimohara, K.: Molecular Computation and Evolutionary Wetware: A Cutting-edge Technology for Artificial Life and Nanobiotechnologies. IEEE Transactions on Systems, Man and Cybernetics, Part C 37(3), 325–336 (2007)

[2] Tanev, I., Brozozowski, M., Shimohara, K.: Evolution, Generality and Robustness of Emerged Surrounding Behavior in Continuous Predators-Prey Pursuit Problem. Genetic Programming and Evolvable Machines 6(3), 301–318 (2005)

[3] Tanve, I., Shimohara, K.: Evolution of Human Competitive Driving Agent Operating a Scale Model for a Car. In: SICE Annual Conf. 2007, pp. 1582–1587 (2007)

[4] Tanev, I.: DOM/XML-based portable genetic representation of the morphology, behavior and communication abilities of evolvable agents. Artificial Life and Robotics 8(1), 52–56 (2004)

[5] Hemmi, H., Maeshiro, T., Shimohara, K.: New Computing System Architecture for Simulation of Biological Signal Transduction Networks. Frontiers of Computational Science, pp. 177–180. Springer, Heidelberg (2007)

[6] Maeshiro, T., Hemmi, H., Shimohara, K.: Ultra-Fast Genome Wide Simulation of Biological Signal Transduction Networks: Starpack. Frontiers of Computational Science, pp. 243–246. Springer, Heidelberg (2007)

[7] Maeshiro, T., Nakayama, S., Hemmi, H., Shimohara, K.: An evolutionary system for the prediction of gen regulatory networks in biological cells. In: SICE Annual Conf. 2007, pp. 1577–1581 (2007)

[8] Takahashi, K., et al.: Bioinformatics 19, 1727–1729 (2003)

[9] Fruchterman, T.M.J., Reingold, E.M.: Software - Practice and Experience 21, 1129–1164 (1991)

# Dynamical Effect on Tick-Wise Price Predictions

Mieko Tanaka-Yamawaki[1] and Keita Awaji[2]

[1] Department of Information and Knowledge Engineering, Tottori University,
4-101 Koyamacho-Minami, Tottori, 680-8550 Japan
[2] Fujitsu Ten, Inc., Kobe, Hyogo, 652-8510 Japan
mieko@ike.tottori-u.ac.jp

**Abstract.** We attempt to incorporate dynamical effect to the tick-wise price prediction, in order to improve performance of the autonomous price prediction generator we construct. Our assumption is that such dynamical effect is carried by local parameters including derivatives of the price (velocities and accelerations) in addition to the price itself, and two dimensionless parameters constructed by using the derivatives. For this purpose, we add a new procedure to the prediction generator that computes derivatives of the data from each segment of the price time series and label that segment by those dynamical parameters. We show in this paper that this dynamical version of the price generator indeed performs better compared to the old version.

**Keywords:** Dynamical Pattern Classifier, Price Prediction Generator, Tick-wise Price, Quadratic Least Square Estimate (QLSE), Stylized facts.

## 1 Introduction

The mechanism of price formation is a fascinating challenge whose answer is still hidden behind thick layers of undiscovered facts. We approach this question by investigating the real tick-wise price time series data provided by foreign exchange markets. Although financial time series are usually assumed to be the random walk, there are plenty of evidences to support the idea that the tick-wise financial data show distinct deviations from the random walk, such as fat-tail, volatility clustering etc., sometimes called as “stylized facts” [1].

However, there are objections against those evidences, attributing them to the effect of mistreatment of unstable time series [2] . Is it a right direction of thought, however, to subtract various sources of noise from financial time series and look for “pure” time series data? Such subtractions may wash out the essence of the price formation mechanism. We would like to try another direction of research toward the discovery of the truth, by taking the “contaminated” data and find the rule inside them.

Recently, we have constructed a price prediction generator that autonomously interpret the past tick-wise data to predict the price trend at a few ticks (one minute or around) ahead of the predicting point with accuracy as high as 70%. The system is based on the evolutional algorithm choosing the best combination of popular technical indicators case by case for the past data of foreign exchange market for several years

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 442–449, 2008.

from 1995 to 2000 [3]. This fact encourages us to believe that the tick-wise motion is indeed predictable.

In this paper, we attempt to improve the performance of our generator by adding a new element, the dynamical effect. The technical indicators that we used in Ref.[3,4] are made from the price data such as moving averages, but did not explicitly include the speed or acceleration of the price changes. Here we incorporate those parameters by computing the local derivatives of the prices from the segments of the time series, by applying quadratic least square estimate (QLSE) algorithm.

The rest of the paper is constructed as follows. In Section 2, we show the basic idea on which the generator extracts dynamical information by scanning data. Then in Section 3, the design of our new price predictor is presented with the way to incorporate dynamical parameters into the system more concretely. The result of applying the new method on the tick-wise price data of USD/JPY exchange rates from 1996 to 2000 is presented and compared to the results given by the old method of Ref.[3] in Section 4 Then we conclude the paper in Section 5.

## 2 Elements of Dynamical Pattern Classifier

### 2.1 Dynamical Price Prediction

It is now widely known that tick-wise prices have strong correlation between adjacent times, which escapes from the standard "random walk" assumption of the traditional financial engineering. In particular, the price of the immediate future at 1-tick ahead is repulsive [5], which hints us to construct a tick-wise price generator [6].

We have proposed in Ref. [3,4], a price prediction generator by using the best combination of technical indicators. This method essentially uses the deviation of the current price from a certain average values of the prices over neighboring time steps and does not consider the velocity or the acceleration of the prices. We attempt to incorporate those dynamical properties as a new set of indicators and utilize their patterns in order to predict the price range at near future.

### 2.2 Quadratic Least Square Estimate (QLSE)

We use the quadratic least square method (QLSE) for each segment of the price time series of length n. Defining the time t within each segment to be $0 \leq t < n$ , and the price at the time $t$ to be $p(t)$, we extract the initial price, $p(0)$ and the initial velocity (of the price) $p'(0)$, and the acceleration $p''(0)$ as the dynamical parameters that represent each segment.

For the linear LSE, the initial price and the initial velocity for each segment are defined as A and B, respectively in the following equation:

$$p(t) = A + Bt \tag{1}$$

They are computed by solving the following coupled linear equation.

$$\begin{bmatrix} 1 & \bar{t} \\ \bar{t} & \overline{t^2} \end{bmatrix} \begin{bmatrix} A \\ B \end{bmatrix} = \begin{bmatrix} \bar{p} \\ \overline{tp} \end{bmatrix} \tag{2}$$

For the QLSE, the initial price, velocity and acceleration (of the price) for each segment are defines as $\alpha$ , $\beta$ , $\gamma$ , respectively in the following equation:

$$p(t) = \alpha + \beta t + \frac{1}{2}\gamma t^2 \tag{3}$$

Those parameters are obtained by solving the following coupled equation:

$$\begin{bmatrix} 1 & \bar{t} & \overline{t^2} \\ \bar{t} & \overline{t^2} & \overline{t^3} \\ \overline{t^2} & \overline{t^3} & \overline{t^4} \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \\ \gamma/2 \end{bmatrix} = \begin{bmatrix} \bar{p} \\ \overline{tp} \\ \overline{t^2 p} \end{bmatrix} \tag{4}$$

Note that A and B that correspond to the initial price and the initial velocity obtained in the linear LSE are slightly different in value to $\alpha$ , $\beta$ obtained in the QLSE. The third dynamical parameter $\gamma$ represents the acceleration for the segment.

### 2.3 Dimensionless Parameters

In this section, we define dimensionless dynamical parameters by using the velocity and acceleration parameters obtained in the last section. $A$, $B$ , or $\alpha$ , $\beta$ , $\gamma$. Dimensionless parameters do not depend on the choice of the units of the price and the time thus expected to have a universal value.

There is one such parameter made of the variable (in our case, the price), its time derivative (velocity), and the second derivative (acceleration), which is called as F-number. The name comes from the similarity to the F-number often used in fluid mechanics representing the ratio of inertia of the fluid over the gravity defined as follows [7].

$$F = \frac{v^2}{x \cdot a} \tag{5}$$

In our case, we use $\alpha$ , $\beta$ , $\gamma$ in place of $x$ , $v$ , $a$ .

$$F = \frac{\beta^2}{\alpha \cdot \gamma} \tag{6}$$

We also define another dimensionless number, $T$ , following Ref. [3] that consists of time interval ( $t$ ), distance ( $L$ ), and velocity ( $v$ ) as follows:

$$T = \frac{v \cdot t}{L} \tag{7}$$

This parameter is named as $T$ since the major effect comes from the time interval ($t$). Unlike F, this parameter $T$ essentially depends on the time scale, since the time interval ($t$) has to be determined as a pre-determined universal parameter. We assume this t be the length of the segment ($n$). Thus the second dimensionless parameter $T$ is defines by means of $n$, $B$, and $A$ for $t$, $v$, $L$, respectively.

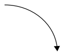
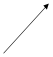
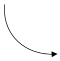
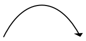
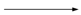
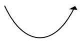
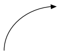
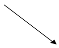
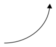
$$T = \frac{n \cdot B}{A} \tag{8}$$

From now on, we consider the effect of parametrization in terms of the above defined dynamical parameters, $F$ ,$T$, $\alpha$, $\beta$, $\gamma$.

### 2.4 Dynamical Patterns

We use two dimensionless dynamical parameters, $T$ and $F$, for the pattern classification. However, tick-wise changes involve extremely short time interval ranging from less than a minute to a few minutes and the price changes are usually very small, 0.01-0.1. For this purpose, we divide the range of velocity $v$ into 3 regions by two threshold points $v_{down}$<0 and $v_{up}$>0, and do the same for the range of acceleration $a$ by two threshold points $a_{down}$ <0 and $a_{up}$ >0. By doing this, the 2 dimensional space of $v$ and $a$ divided into 9 regions, as shown in Table 1.

**Table 1.** Dynamical pattern classification by means of $v$ and $a$

| | $a < a_{down} < 0$ | $a_{down} < a < a_{up}$ | $0 < a_{up} < a$ |
|---|---|---|---|
| $v < v_{down} < 0$ |  |  |  |
| $v_{down} < v < v_{up}$ |  |  |  |
| $0 < v_{up} < v$ |  |  |  |

In the scheme on Table 1, there are 9 patterns. The velocity $v$ indicates the up/down of the price within each segment, while the acceleration $a$ indicates the up/down of the derivative of the price. The dynamical parameter $B$ and $\gamma$ can be regarded to correspond to the velocity $v$ and $a$ in Table 1. However, the dimensionless parameters $F$ and $T$ requires not only $B$ and $\gamma$ but $A$, $\alpha$, $\beta$ and the corresponding tabulation by means of $F$ and $T$ are somewhat more complicated. By using the patterns as illustrated, we can handle more delicate pattern classification compared to the standard ways depending solely on price up/down patterns.

## 3 Intra-day Forecast by Means of Dynamical Pattern Classifier

### 3.1 Job Flow of the Dynamical Price Prediction Generator

The basic structure of the job flow as follows, which is similar to the generator that we proposed before in the study of technical indicator combinations [3].

① Set up the parameters in the prescribed range given by
a) Pattern length, n: 3-ticks< $n$ <30-ticks
b) Term of Prediction Experiment, L: data length of one day (8000-9000ticks for 2000 exchange rate of USD/JPY.
c) Predicted Range, R=1-10 ticks ahead of the point of computation

② Generate a prediction strategy based on the dynamical parameters

③ Compute the dynamical parameters ‘A, B,α,β,γ, F, T from the data pieces

④ Extract patterns by means of the dynamical parameters

⑤ Use the prediction strategy and make a prediction

⑥ Repeat ③-⑤ for L times

Compute the hitting rate as a ratio of correctly predicted direction of move divided by the total number of predictions L, and evaluate the strategies. If there are more than 1000 strategies, select the best 1000 strategies according to the performance in the evaluated hitting rates and perform the same genetic operation as in Ref. [3]. Namely, we sort the current strategies according to the order of performance and send the top 10% genes and their 9 different mutants to the next generation.

⑦ Repeat ⑥-⑦ for all the data applied for prediction experiment.

### 3.2 Generation of Prediction Strategy

The local velocity $B$ and the local acceleration $\gamma$ obtained by QLSE for each data segment are used as the dynamical parameters for pattern classification. The 9 patterns correspond to the 3 by 3 matrix of $B$ and $\gamma$ divided by $B_{up}$, $B_{down}$, $\gamma_{up}$, and $\gamma_{down}$, respectively. The prediction strategy is a gene corresponding to the leaves of the tree of order 9, corresponding to the 9 patterns of dynamical variables at each time step.

A typical example of the strategy generation for n=1 is illustrated in Fig.1, where the leaves of the tree of depth n=1 have the prediction learned from the past data. The price prediction generator has a set of strategy corresponding to all the possible event history of n steps. Under the prediction mode, the prediction generator recognizes the pattern just occurred in the last n steps and answer the strategy written in the leaf at the end of the corresponding path in the event tree. For example, if the event just

**Table 2.** The nine patterns of $B$ and $\gamma$

| | $\gamma < \gamma_{down} < 0$ | $\gamma_{down}<\gamma<\gamma_{up}$ | $0<\gamma_{up}<\gamma$ |
|---|---|---|---|
| $B < B_{down} < 0$ | ① | ② | ③ |
| $B_{down}<B<B_{up}$ | ④ | ⑤ | ⑥ |
| $0< B_{up} < B$ | ⑦ | ⑧ | ⑨ |

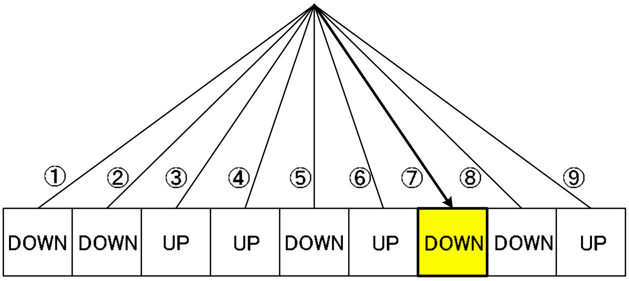


**Fig. 1.** String representation of strategy as leaves of an event tree of order 9, corresponding to the minimum history length (n=1) in the classification of dynamical parameters shown in Table 2

occurred is the seventh pattern ($\gamma < \gamma_{down} < 0$ and $0 < B_{up} < B$), the prediction is "DOWN". The results of 1000 strategies are tested and the rates of correct prediction are recorded for each strategy. At the end of one experiment (scanning of the data segment of length L), evaluation and selection of the strategies are performed to evolve the set of the strategies for the next generation.

For the case of 9 patterns, there exist 512 possible strategies for the tree of minimum depth, n=1, and $2^{81}$ possible strategies for the tree of depth 2 corresponding to the memory length n=2. We set the maximum number of strategies to be 1000 for the sake of computational time. We use all the strategies if the total number does not exceed 1000. On the other hand, if the number of strategy exceeds 1000, we follow the same evolutional technique as we used in Ref. [3] to select the 1000 best strategies. The results reported in the following sections are the case of n=1, thus no evolutional algorithm works and all the possible strategies are examined.

### 3.3 Evaluation of Hitting Rates and Evolution of Strategies

After repeating the prediction process for L steps, the system terminates the experiment in order to evaluate the strategies by the rate of correct predictions. At this point, the best strategy and its hitting rate, as well as the average hitting rate of 10 best strategies are recorded and new strategies are prepared for the next generation by means of an evolutional algorithm. The hitting rate is defines as follows.

$$\text{Hitting rate} = [\text{success}]/[\text{success} + \text{failure}] \qquad (9)$$

Here “success” means the events for which the actual direction of move and the predicted direction of move matched, and “failure” means the opposite. The events for which the price did not move are excluded from the denominator.

# 4 Results

## 4.1 Results of Prediction Experiment

In our experiment, we have used one pattern as a conditional part of the conditional probability thus no need of evolutional mechanism.

We show the hitting rates of the best strategies of this new generator, compared with the result of old generator in Ref.[1] in Fig.2 applied on the data of foreign exchange rate USD/JPY from 1996 to 2000. The upper two line are the result of predicting 1 tick ahead of the predicting time and the lower two lines are the result of predicting 10 ticks ahead. Both cases show that the new dynamical version outperforms the old version of our predictor.

The picture of the hitting rate as a function of the predicted point ranging from 1-10 tick is shown in Fig.3. From Fig.2-3, we observe that our new result always performs better by 0.5-2%.

Another factor to concern is the length of the pattern, n. The old result obtained by using the evolutional method in Refs. [3, 4, 6, 8, 9] the history length H in the range of 1-5. Since the best result was obtained for H=3, the number of possible strategies are $3^{27}$. From now on, the ‘past’ result indicates the best hitting rate obtained for H=3 strategies. In the current analysis of using the dynamical parameters, similar effects have been observed. The result of computing the dynamical parameters over 5 ticks by means of our new method performs worse than the case of using 3 ticks in the same method, they still outperform the results of ‘past’ method in more than a half of

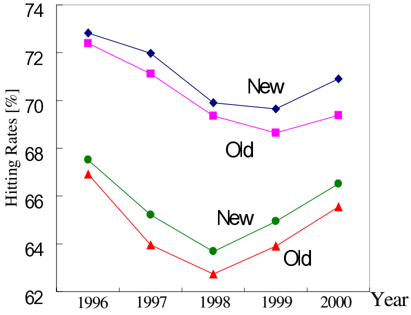


**Fig. 2.** Comparison of the hitting rates [%] of the new/old method at 1-tick ahead (upper) /10 ticks ahead (lower) on the tick data of USD/JPY, from the year 1996 to 2000

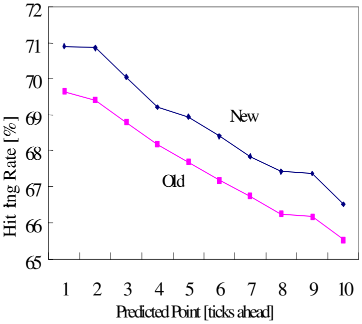


**Fig. 3.** Hitting rates [%] of the new/old method vs. the range of the predicted point, 1 tick-10. ticks measured forUSD/JPY2000.

the entire data. However, the new method using more than 10 ticks of history considerably ill-performs and the hitting rate further goes down as we increase the range of history and becomes random after the range reaches 20-30 ticks.

## 5 Conclusion

In this paper, we applied a novel version of our prediction generator that reads and uses the patterns of dimensionless dynamical parameters together with the local velocities and accelerations on the tick-wise price changes, reflecting the correlation between tick-wise prices. By doing this, we have successfully improved the rate of correctly predicting the future direction of the price range by 1-2% based on the patterns of 3 ticks to 5 ticks in the past. The result of predicting 10 ticks ahead also shows improvement in comparison to our past result in Ref.[3,4] and Ref.[6].

However, the use of past patterns of 10 ticks or older considerably lowers the rate of correct prediction on the up/down trends of the price range. The use of past patterns of 20 ticks or older turns out meaningless since the prediction based on those information shows random series of up/down trends.

Based on this fact, we conclude that the meaningful size of segments lies between 3 to 6 ticks, and not only the sign of the price changes but the dynamical patterns including velocity and acceleration play important effect on the prediction.

Although it is still unclear whether dimensionless parameters are effective on the improvement of the rate of correct prediction of the future price range, the best parametrization have been the 9 patterns of the 3 velocity times the 3 acceleration.

## References

[1] Mategna, R.N., Stanley, H.E.: An Introduction to Econophysics: Correlations and Complexity in Finance. Cambridge Univ. Press, Cambridge (2000)

[2] Takayasu, H.(ed.): Empirical Science of Financial Fluctuantions. Springer, Heidelberg (2001)

[3] Tanaka-Yamawaki, M., Tokuoka, S.: Adaptive Use of Technical Indicators for the Prediction of High-frequency Price Fluctuations. In: Apolloni, B., Howlett, R.J., Jain, L. (eds.) KES 2007, Part II. LNCS (LNAI), vol. 4693, pp. 597–603. Springer, Heidelberg (2007)

[4] Tanaka-Yamawaki, M., Tokuoka, S.: Adaptive Use of Technical Indicators for the Prediction of Intra-day Stock Prices. Physica A 383, 125–133 (2007)

[5] Tanaka-Yamawaki, M.: Stability of Markovian Structure Observed in High Frequency Foreign Exchange Data. Ann. Inst. Statist. Math. 55, 437–446 (2003)

[6] Tanaka-Yamawaki, M., Motoyama, T.: Predicting the Tick-wise Price Fluctuations by Means of Evolutionary Computation. In: IEEE-CEC 2004, pp. 955–958 (2004)

[7] Andersen, J.V., Gluzman, S., Sornette, D.: Fundamental Framework for Technical Analysis. European Physical Journal B 14, 579–601 (1999)

[8] Tanaka-Yamawaki, M.: On the Predictability of High-Frequency Financial Time Series. In: Palade, V., Howlett, R.J., Jain, L. (eds.) KES 2003. LNCS, vol. 2773, pp. 1100–1108. Springer, Heidelberg (2003)

[9] Tanaka-Yamawaki, M.: Tick-wise Predictions of Foreign Exchange Rates. In: Negoita, M., Howlett, R.J., Jain, L.C. (eds.) KES 2004. LNCS (LNAI), vol. 3213, pp. 449–454. Springer, Heidelberg (2004)

# Time-Series Models of the EEG Wearing Overcorrected Glasses

Yohei Tomita[1], Shin-ichi Ito[1], Yasue Mitsukura[1],
Minoru Fukumi[2], and Taketoshi Suzuki[3]

[1] Tokyo University of Agriculture and Technology,
2-24-16, Naka, Koganei, Tokyo, Japan
50008401213@st.tuat.ac.jp,{ito_s,mitsu_e}@cc.tuat.ac.jp
[2] University of Tokushima,
2-1, Minami-Josanjima, Tokushima, Japan
fukumi@is.tokushima-u.ac.jp
[3] Suzuki Eye Clinic Kichikouji,
16, kichikoji, Mizusawa, Osyu, Japan
suzu-kichikoji@world.ocn.ne.jp

**Abstract.** Most people do not notice the overcorrected glasses in daily life. The overcorrected glasses have harmful effects on the eye. Then, these effects are thought to have harmful effects on the brain, too. Therefore, to reveal the effects on the brain by the overcorrection, we analyze electroencephalogram (EEG). At the experiment, the subject played the PC game for 30 minutes (techno-stress) with overcorrected glasses. As the results for time-series analysis and average-variance analysis, the differences of the EEG feature between the correction and the overcorrection are confirmed.

**Keywords:** EEG, the overcorrection, techno-stress, time-series analysis.

## 1 Introduction

IT equipments such as personal computers and game machines have been increasing dramatically. However, the actual condition of techno-stress is little known. Techno-stress causes occupational fatigue, stiff shoulder, neck pain, ocular pain and tired eye [1], [2], [3]. Especially, most people have been suffered from eye damage because techno-stress has harmful effects on the eye. These harmful effects cause blurred vision, ocular pain, tears, nauseous and so on. Furthermore, most people does not notice their eyeglasses overcorrected in daily life and that enhances techno-stress [4].

Harmful effects from techno-stress are not appearance only in the eye but the brain because techno-stress causes mental stress. Mental stress is for the most part generated unconsciously, and feelings often follow on physiological changes in the brain [5], [6]. For these reasons, we use the EEG for revealing harmful effects on the brain by the overcorrection. It is a recording of brain activity and the pattern of activity changes with the level of brain activity.

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 450–457, 2008.

In this study, we reveal the harmful effects by overcorrection using the EEG. At the experiment, the subjects wear the overcorrected glasses and play PC game. Then, we make the EEG measurement before starting the game, during the gaming, and after the gaming. The analysis methods are time-series analysis and average-variance analysis.

## 2 Experimental Procedure

### 2.1 Overcorrection and Techno-stress

Eyeglasses for the subjects are chosen by an ophthalmologist (Coauthor Taketoshi Suzuki). One is corrected eyeglasses, and the other is overcorrected eyeglasses. The overcorrected eyeglasses are set by -1.5D. It is practically unnoticeable load so that ordinary people do not notice that their glasses are overcorrected. In this study, we are clarifying effects on the brain by the overcorrected eyeglasses, compared with the corrected eyeglasses. Furthermore, to give the subjects techno-stress, the subject played PC game for 30 minutes with a pair of glasses and the subjects wear an electroencephalograph (Fig.1) during the gaming.

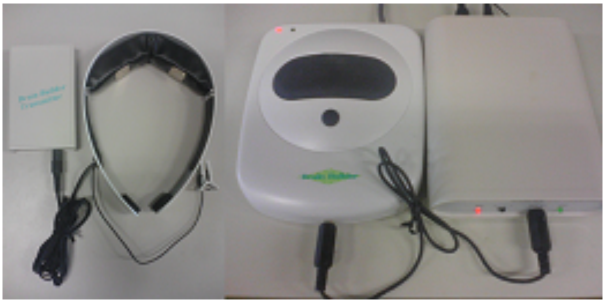


**Fig. 1.** Brain Builder

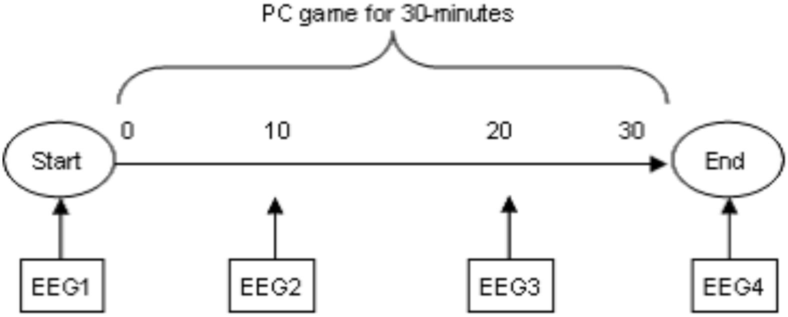


**Fig. 2.** The flow of experimental procedure

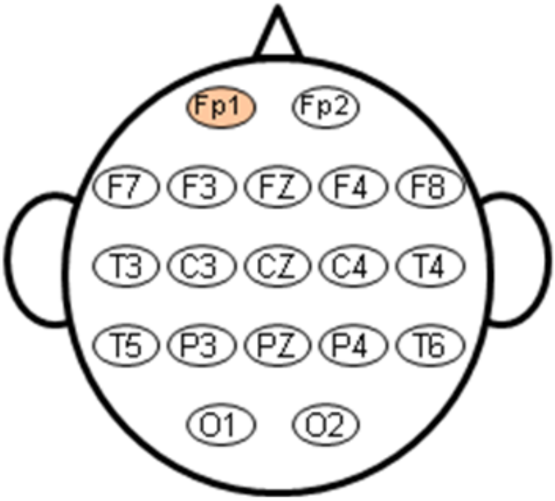


**Fig. 3.** The international 10-20 system

The flow of the experimental procedure is shown in Fig.2. Before playing the game, the subjects are relaxed with their eyes closed and the EEG is measured. Then, during playing the game, the EEG measurement is carried out twice. Finally, after playing the game, the EEG measurement is carried out again. At all the measurements, the subjects keep relaxed on the chair with their eyes closed.

## 2.2 Measurement of the EEG

We use the simple electroencephalograph of the band type made by Brain Function Research and Development Center. This electroencephalograph measures the EEG activity at FP1 in the international 10-20 system shown in Fig.3. Measurement methodology is referential recording: reference electrode is at the left ear lobe and exploring electrode is at FP1 [7]. This electroencephalograph can obtain an one-channel EEG data through the serial port. This data is sent to the computer every second through the serial port for the Fast Fourier Transform (FFT) calculation (Sampling frequency is 128$Hz$). The range of frequencies is 4-22$Hz$ at 1$Hz$ intervals. In addition, we use 5 frequency bands ($\theta$ wave: 4-6$Hz$, slow$\alpha$ wave: 7-8$Hz$, mid$\alpha$ wave : 9-11$Hz$, fast$\alpha$ wave: 12-14$Hz$, $\beta$ wave: 15-22$Hz$).

The measurement time is 60 seconds per one measurement and the number of measurement is 4 times (before the experiment, in 10 minutes, in 20 minutes, after the experiment). Although the subjects' eyes are opened to look at a computer screen during playing the game, the measurement is started with their eyes closed.

## 3 Proposed Method

Extracting the EEG features is absolutely imperative for analysis. Extracted features lead to find the effects on the brain by the overcorrection. First of all, the frequency bands features of the correction and overcorrection is extracted, using the 1 measurement average of frequency band. Secondly, time-series models of the EEG are extracted by the time-series analysis. Finally, by the average-variance analysis, the features of time-series models are analyzed in more details.

At the time-series analysis, autocorrelation coefficient is used for extracting time-series models. As already mentioned, overcorrection causes the tiredness of the brain and we assume that the tiredness of the brain increased time-series instability of the EEG. Therefore, autocorrelation can be used to extract instability of the EEG. The autocorrelation is given by the following equations:

$$R_k = \frac{Cov(x_n, x_{n-k})}{\sqrt{Var(x_n)Var(x_{n-k})}} \ . \qquad (1)$$

$R$ is autocorrelation to time-series. Sampling $(n)$ is 128 orders (1 seconds) and the time lag $(k)$ is 128 points (1 seconds). Therefore, time length of $R_k$ is 2 seconds. When a length of the EEG data is 2 seconds (Sampling frequency is $128Hz$), 129 different $R_k$ are obtained $(0 \leq k \leq 128)$.

Average-variance analysis is the method to know average and variance $(Ave(y_N), Var(y_N))$ of observed data. Then, this analysis is used for additional analysis on time-series models extracted by the time-series analysis.

## 4 Results

We chose 2 people (SubjectA and SubjectB) for the subjects. Results of investigation are at the following. First of all, we analyzed the features of frequency band. Secondly, we extracted time-series models. Finally, we analyzed the features of 10 $Hz$ to analyze in more detail.

### 4.1 Average of Frequency Band

In order to analyze the EEG change, averages of the Fourier spectra on each frequency band of the EEG were used shown in Table.1 and Table.2.

At the Table.1, the spectra averages of each frequency band as in case of the overcorrection compared with in case of the correction are: $\theta$ is +1.90, slow$\alpha$ is +2.77, mid$\alpha$ is +0.66, fast$\alpha$ is +1.62, and $\beta$ is +0.83. Then, at the Table.2, the spectra averages of each frequency band are: $\theta$ is +2.24, slow$\alpha$ is +3.88, mid$\alpha$ is +2.33, fast$\alpha$ is +1.43, and $\beta$ is +2.14. Therefore, it is inferred from this result that the EEG of the overcorrection tends to higher than that of the correction.

### 4.2 Time-Series Models

There are 2 main time-series structures by the periodicity of autocorrelation (Fig.4 and Fig.5). At these figures, the vertical axis is autocorrelation coefficient

**Table 1.** Average of each frequency band (SubjectA)

| | $\theta$ | slow$\alpha$ | mid$\alpha$ | fast$\alpha$ | $\beta$ |
|---|---|---|---|---|---|
| Correction | 9.68 | 14.17 | 11.73 | 9.36 | 7.56 |
| Overcorrection | 11.58 | 16.95 | 12.39 | 10.98 | 8.39 |

**Table 2.** Average of each frequency band (SubjectB)

| | $\theta$ | slow$\alpha$ | mid$\alpha$ | fast$\alpha$ | $\beta$ |
|---|---|---|---|---|---|
| Correction | 6.47 | 9.36 | 8.53 | 7.48 | 6.41 |
| Overcorrection | 8.71 | 13.51 | 10.86 | 8.91 | 8.55 |

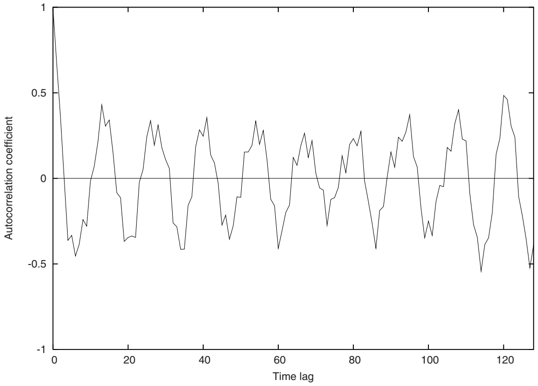


**Fig. 4.** Time-series model 1 (autocorrelation)

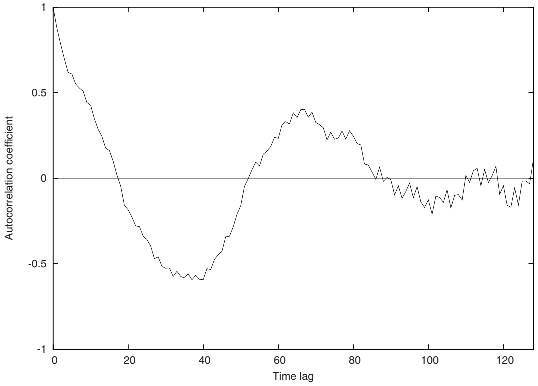


**Fig. 5.** Time-series model 2 (autocorrelation)

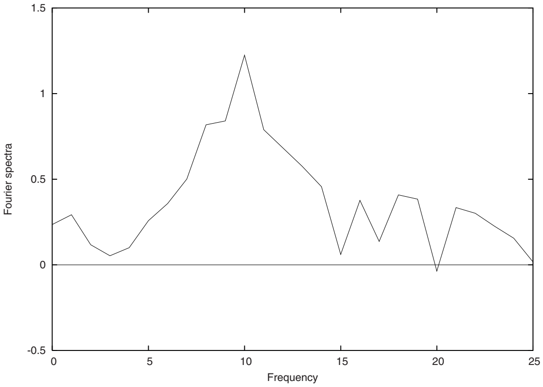


**Fig. 6.** Time-series model 1 (frequency)

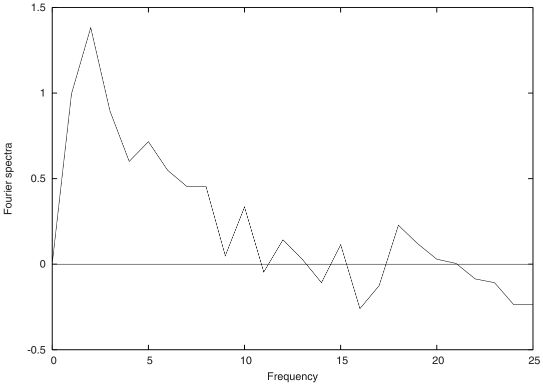


**Fig. 7.** Time-series model 2 (frequency)

and the abscissa axis is lag ($k$). Then, for extracting frequency features of these structures, Fourier transform is computed (Fig.6 and Fig.7). At these figures, the vertical axis is the Fourier spectra on logarithmic display and the abscissa axis is frequency.

Fig.4 and Fig.6 indicate the structure 1 which means that the wave of about $10Hz$ constantly appears. Peak coefficient was at $10Hz$. On the contrary, Fig.5 and Fig.7 indicate the structure 2 which means that the amplitude and the frequency decrease with an increasing lag. Furthermore, 1/f fluctuation was observed in the Fig.7. These structures have different occurrence rate. The structure 1 tends to appear with the corrected glasses. The structure 2 tends to appear with the overcorrected glasses and as time advances.

### 4.3 Average and Variance of $10Hz$

As the time-series analysis results, near $10Hz$ appear at an early stage of the techno-stress especially with the overcorrected glasses. Table.3 and Fig.4

**Table 3.** Average of 10 $Hz$

| | SubjectA | SubjectB |
|---|---|---|
| Correction | 36.51 | 28.88 |
| Overcorrection | 66.06 | 36.79 |

**Table 4.** Variance of 10 $Hz$

| | SubjectA | SubjectB |
|---|---|---|
| Correction | 11.32 | 8.57 |
| Overcorrection | 12.08 | 11.71 |

indicate the features of 10$Hz$ because the features of this stable wave clearly appear on 10$Hz$. Although there are 4 EEG data thorough 1 trial, these parameters are calculated among all 4 data. At these tables, these parameters at the overcorrection increased.

## 5 Discussions

From the results, it is observed that there are different EEG features between the correction and the overcorrection (Table.5). When the subject wore the corrected glasses, the spectrum of the EEG is low compared with the overcorrected. Especially, this feature strongly appears at 10$Hz$. Furthermore, the structure 1 indicate stable wave whose frequency is near 10$Hz$. Therefore, it is thought that the EEG is generated stably similar characteristic of the $\alpha$ wave. On the contrary, there are quite different features as in case of the overcorrection. The spectrum of the EEG is high and the structure 2 is founded. The structure 2 is a damped oscillatory wave, so that the EEG is generated unstably. This reason is that the overcorrected glasses give load to the subjects' brain.

From the previous research, it is known that the $\alpha$ wave significantly appears during the resting state. In this study, it is confirmed that the $\alpha$ wave constantly appeared in case of the correction, not the overcorrection. Therefore, the correction could cause the subjects resting state, and the overcorrection could not cause the subjects resting state. As already mentioned, it is difficult to notice overcorrection daily life. However, from this study, it is obvious that the

**Table 5.** Overcorrection and the EEG

| | Correction | Overcorrection |
|---|---|---|
| Average of frequency bands | Decrease | Increase |
| Time-series models | Structure 1 | Structure 2 |
| Average and Variance of 10Hz | Decrease | Increase |

overcorrection effects on the brain. At the future works, we quantify the occurrence rate of time-series models, in order to evaluate objectively. Furthermore, it is important to analyze the correlation between the structure 2 and the brain.

## 6 Conclusions

In this paper, we research the effects on the brain by the overcorrection. At the experiment, the subject is given the techno-stress by playing the PC game. For comparison between corrected glasses and overcorrected glasses, the subject wore both glasses alternately. Then, according to time-series analysis and average-variance analysis, different features of the EEG are confirmed between the correction and overcorrection. As the results, it is confirmed that the $\alpha$ wave constantly appeared in case of the correction, not the overcorrection.

## References

1. Ye, Z., Abe, Y., Kusano, Y., Takamura, N., Eida, K., Tai-ichiro, Takemotom, Aoyagi, K.: The Influence of Visual Display Terminal Use on the Physical and Mental Conditions of Administrative Staff in Japan. Journal of Physiological Anthropology 26(2), 69–73 (2007)
2. Iwanaga, K., Liu, X.X., Shimomura, Y., Katsuura, T.: Approach to Human Adaptability to Stresses of City Life. Journal of Physiological Anthropology and Applied Human Science 24(4), 357–361 (2005)
3. Kasuga, N.: Study on Technostress Syndrome (Report 1). Japanese Journal of Psychosomatic Medicine 32(5), 383–390 (1992)
4. Tokoro, T., Kabe, S.: Treatment of the myopia and the changes in optical components. Report 2. Full-or under-correction of myopia by glasses Acta. Soc. Ophthalmol. Jpn. 69, 140–144 (1965)
5. Matsunami, K., Homma, S., Han, X.Y., Jiang, Y.F.: Generator Sources of EEG Large Waves Elicited by Mental Stress of Memory Recall or Mental Calculation. Japanese Journal of Physiology 51, 621–624 (2001)
6. Sutton, S.K., Davidson, R.J.: Menstrual Cycle Effects on Performance of Mental Arithmetic Task. Journal of Physiological Anthropology and Applied Human Science 21(6), 285–290 (2002)
7. Kasamatsu, K., Suzuki, S., Anse, M., Funada, M.F., Idogawa, K., Ninomija, S.P.: Prefrontal brain electrical asymmetry predicts the evaluation of affective stimuli. Neuropsychologia 38, 1723–1733 (2000)

# Feature Extraction System for Age Estimation

Hironobu Fukai[1], Hironori Takimoto[2], Yasue Mitsukura[1], and Minoru Fukumi[3]

[1] Graduate School of Bio-Applications & Systems Engineering,
Tokyo University of Agriculture and Technology,
2-24-16, Naka-cho, Koganei, Tokyo, 184-8588, Japan
50007701291@st.tuat.ac.jp,
mitsu_e@cc.tuat.ac.jp

[2] Department of Control Engineering, Sasebo National College of Technology,
1-1, Okishin-cho, Sasebo, Nagasaki, 857-1193, Japan
takimoto@post.cc.sasebo.ac.jp

[3] Faculty of Engineering, The University of Tokushima,
2-1, Minami-Josanjima-cho, Tokushima, 770-8506, Japan
fukumi@is.tokushima-u.ac.jp

**Abstract.** In this paper, we propose the novel age estimation system with the real-coded genetic algorithm (RGA) and the neural network (NN). The age is one of important information in our living. There are a lot of studies on age estimation by the computer. However, the conventional method of the age estimation, the most of them are the studies intended for an actual age. Therefore, we pay attention to the mechanism of human age perception. The apparent age feature is extracted by the fourier transform, and the important spectrum for the age perception are selected by the RGA. The age is estimated by the 3 layered NN. It is considered that it can extract important age feature using the RGA and it can analyze the important feature area. In addition, proposed method extracts the age feature at each age. In order to show the effectiveness of the proposed method, we show the simulation examples. From the simulation results, we can confirm that the proposed method works well.

**Keywords:** age estimation, neural network (NN), real-coded genetic algorithm(RGA).

## 1 Introduction

We can easily estimate a person's age from a face image. Moreover, we can take a smooth and flexible correspondence by estimation the age. For this reason, it is considered that the age is one of the most important information in our living. Therefore, the age estimation method by face image was widely studied [1]-[11].

Todd *et. al.* indicate that contour of skull are approximated by cardioid transform [1,2]. Yamaguchi *et. al.* show the difference of the feature of adult and child's faces was overall information like the length of the face and the ratio of each part [3]. On the other hand, age estimation by computer is performed. Kanno *et. al.* shows that the man was identified by the neural network for four ages

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 458–465, 2008.

(12 years, 15years, 18 years, and 22 years)[4]. Moreover, Y.H.Kwon *et. al.* are reported that the theory has only been implemented to classify input images into one of three age-groups: babies, young adults, and senior adults [5]. The computations are based on cranio-facial development theory and skin wrinkle analysis. Burt *et. al.* studied the age perception that uses the averaged face from 25 to 60 years. Especially, they used texture and shape [6]. Ueki *et. al.* are reported that the age-group classification by the dimension compression [7]. Takimoto *et. al.* proposed the gender and age estimation technique not influenced by the posture change by estimating NN by using several features including the texture features [8].

The conventional methods has a lot of problems for practical use though is expected various applications, for example, age confirmation in vending machine of cigarette, the buyers' investigations in convenience store etc., and so on. In addition, the conventional method of the age estimation, these are the studies intended for an actual age, and there is little study that pays attention to human age perception. S.Mukaida *et. al.* indicate that it is possible to change the age impression by analyzing the skin information, and operating these. This report is shown the relationship between skin information and the apparent age. However, it is a part of human's age perception, and it is shown that it is effective to only correction extent. From this results, the processing process and the characteristic of the human's age estimation is not yet clarified. If we can extract the feature that human uses for potential to the age estimation, it is considered that the sensibility that closes to human beings to the computer and the robot can be given. Moreover, it is effective for the anti aging and cosmetic surgery and so on.

Therefore, we propose apparent age estimation system from the face image based on human's age perception. In the study of age estimation, it is considered that age feature appears face texture information, such as information on wrinkle, pigmented spot, and so on. It is considered that the frequency analysis is effective for the extraction of texture information. However, when we estimate the age, we have changed the seen part by subject's generation. For example, when we estimate a young person, we see skin tone and firmness, and when we estimate a elderly person, we consider wrinkle and pigmented spot. Then, in the proposed method, texture information on the skin is converted as the generation estimates easily to each generation, and the age is estimated. The age feature is extracted by the fourier transform to the face image. The feature of each generation is extracted by the real-coded genetic algorithm (RGA). Moreover, we estimate the age by the neural network (NN). In order to verify the effectiveness of the proposed method, we show the computer simulation based on actual data (HOIP database).

## 2 Preprocessing

It is considered that the face image and normalization and the feature extraction from the face image is necessary to extract the age feature. Moreover, it is necessary to give the apparent age to the face image to estimate the age in

**Table 1.** The detail of the face image database

| | |
|---|---|
| Size | 640×480[pix.] |
| | 24 bit color |
| Gender | 150 images for each |
| Age | 30 images per 5 years old |
| Emotion | neutral |

**Fig. 1.** Example of original images

which it pays attention to human's age perception. This chapter explains these processing.

### 2.1 Face Image Database

The face database is provided from Human and Object Interaction Processing (HOIP) organization in JAPAN [12]. The face images of this database are the people with a wide age group that doesn't sport a pair of glasses. The background and proof were made the same condition for all images. Subject was directed to make the lens of the camera see, and it took a picture with that look of natural (Table 1). Fig.1 shows the example of the original image. 252 people who gave the preprocessing beforehand are used as subject.

### 2.2 Normalization

It is necessary to normalize the face image to the age estimation, because the original image of the data base is not constant the position of the face. Moreover, original images have much unnecessary information.

The face image is normalized based on both eyes. The reason for having used the eye for normalization of face image is as follows. The first, eyes are having

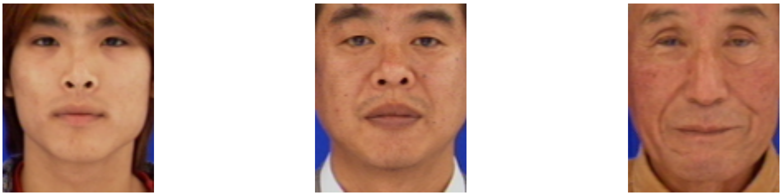
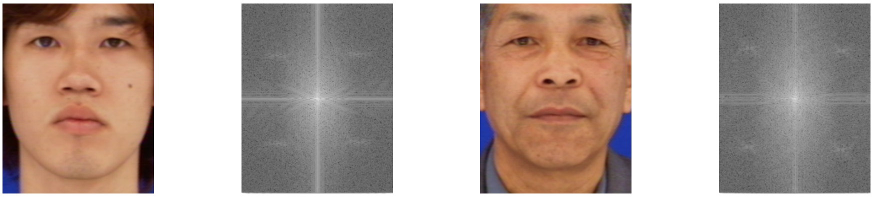


**Fig. 2.** Normalization of the face images



(a)20's man

(b) 60's man

**Fig. 3.** Results of the fourier transform

been easy to perform the normalization about a rotation and a size, compared with other features of face. Next, many researches of extracting the region of an eye are proposed [13,14]. Therefore, to use eye for normalization of face image is efficient.

### 2.3 Feature Extraction from Face Image

It is considered that texture information on the skin is important for human's age. Then, the feature data has been extracted by using the fourier transform for the face image. Fig.3 shows results of the fourier transform. From these results, it is shown that frequency information in the face image has gathered in the low frequency. Moreover, the elderly persons tended to strengthen more than the young person as for high frequencies.

### 2.4 Apparent Age Database

In this paper, the apparent age was given by doing the questionnaire survey to 60 subjects. It is considered that objective apparent age was able to be obtained by using a lot of people. Moreover, the error of age estimation for each age can be reduced by having elected various generations and gender questionnaire subjects.

As the questionnaire method, the subject sees the face images and the apparent age is given. The face image prepares the one arranged at random, and the subject estimate the age from the edge by intuition sequentially. The apparent age was assumed to be a median value of the age that the subject had given. In proposed method, this age is adopted for the teacher data as apparent age.

## 3 Extraction of Age Feature

An important frequency is decided to the age feature by weighting to the frequency feature obtained by the fourier transform. It is considered that the age feature in each generation can be extracted by weighting to each generation. An important frequency is extracted by the RGA.

**Table 2.** The parameters of the NN

| | |
|---|---|
| The number of input layer units | 20250 |
| The number of hidden layer units | 20 |
| The number of output layer units | 4 |
| The number of learning cycles | 1000 |
| Learning coefficient | 0.8 |

**Table 3.** The parameters of the RGA

| | |
|---|---|
| Chromosomes | 81000 |
| Individuals | 100 |
| Generations | 100 |
| Crossover late | 0.9 |
| Mutation rate | 0.01 |
| Elite strategy | use |

### 3.1 Evaluation of Fitness

The fitness function used the error that is estimating the age actually. The genetic algorithm was used as a minimization problem so that the error might become small. The age is estimated by the 3-layered neural network (NN). NN is known to be an especially excellent of the problem related to the pattern recognition. Thus, in this study, the NN is used for age estimation. Details of the NN are described in the next chapter.

## 4 Age Estimation Method

In this study, 3-layered NN is used as an age estimator. The study role used back propagation method. Moreover, to classify the age at 4 generation, the number of output layers is assumed to be 4. The number of units of input layers is assumed to be 20250 that exclude the symmetry part from all frequencies of 40500. The sigmoid function was used for a nonlinear function of NN.

## 5 Computer Simulations

In order to show the effectiveness of the proposed method, we show the simulation examples. In this study, we use the subjects that donft sport a pair of glasses. The teacher data was arbitrarily selected 5 people from each generation, and the test data arbitrarily selected 3 people from each generation. Experimental conditions of NN are shown in Talbe 2, and experimental conditions of GA are shown in Table 3.

First, Fig. 4 shows the convergence of fitness by the GA. Fitness function used the output error of NN that used the test data. In addition, RGA is set for this error margin to become small. 20 people who were teacher data of 5 people in

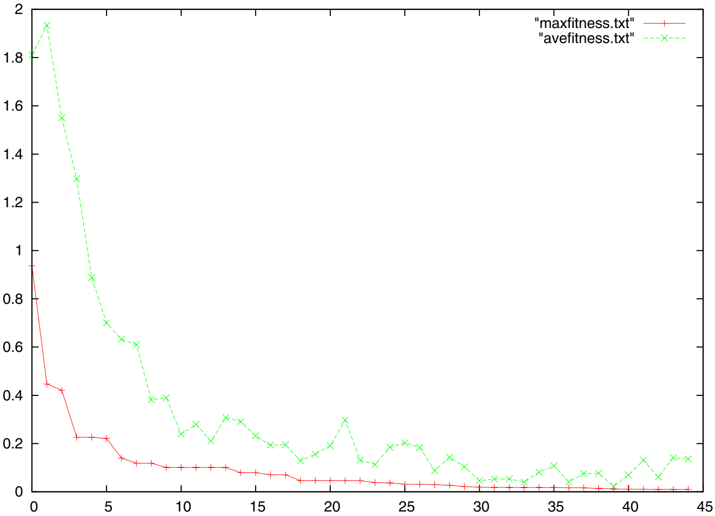


**Fig. 4.** The convergence of fitness

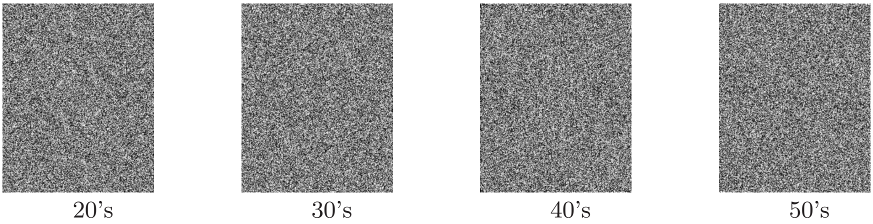


**Fig. 5.** Weight coefficient

each generation were made to study to NN. Furthermore, 12 people who were test data of 3 people in each generation were used for test data. NN was studied until the recognition rate of the study data became 100%. Moreover, the end condition of RGA was assumed that the recognition rate when the test data was used for NN was 100%.

In this technique, the frequency has been weighted in each generation. Fig.5 shows the results of the weight. It displays it in white if weight is near 1, and it displays it in black if weight is near 0. The recognition rate of each generation's test data became 100% by using the weight of Fig.5. By the way, the correlation coefficient of the weight between generations became all less than 0.01 and was not seen the correlation in the feature between generations (Table4). From this result, it was shown that the age feature was quite different depending on the generation. Moreover, this weight of the great difference was not seen as shown in Table5 though requested the average from a low frequency, high frequency, and horizontal frequencies and vertical frequencies. As a result, an important frequency for the age can be said that there is a possibility of existing in all

**Table 4.** The correlation of weight

| | correlation |
|---|---|
| 20's - 30's | -0.0017 |
| 20's - 40's | 0.003 |
| 20's - 50's | -0.0098 |
| 30's - 40's | 0.0099 |
| 30's - 50's | 0.0043 |
| 40's - 50's | -0.0026 |

**Table 5.** The average of frequency

| | Vertical | Horizon | High frequency | Low frequency |
|---|---|---|---|---|
| 20's | 0.496 | 0.519 | 0.505 | 0.504 |
| 30's | 0.473 | 0.531 | 0.506 | 0.501 |
| 40's | 0.497 | 0.494 | 0.502 | 0.5 |
| 50's | 0.506 | 0.538 | 0.497 | 0.5 |

districts about the high frequency, the low frequency, horizontal direction, the vertical direction. Moreover, the average of weight of a horizontal frequency was higher than vertical frequency though it was a minute difference. This is considered that it is shown that horizontal direction influences the age though horizontal gray value information and vertical gray value information is also important. A further verification will be done by increasing the number of test data in the future. Moreover, it is necessary to investigate what features are in weight further.

## 6 Conclusions

In this study, we proposed the novel age estimation method. The feature has been extracted as the age is divided into the delimitation at the age of ten at the fourth generation. The face image was normalized based on both eyes. Moreover, feature at the age are extracted by fourier transform. Furthermore, the apparent age was given by questionnaire survey to various generations. The age feature was extracted to each generation by combining RGA and NN, and the apparent age was estimated. It is considered that the proposed method worked well by the computer simulations.

In the future works, a further verification will be done by increasing the number of test data. Moreover, it is necessary to investigate what features are in weight further.

The face database has obtained the use permission from corporation SOFTPIA JAPAN. It is prohibited to copy, to use, and to distribute without the authorization of the right holder.

## References

1. Todd, J.T., Mark, L.S., Shaw, R.E., Pittenger, J.B.: The perception of human growth. Scientific American Perception 242, 106–114 (1980)
2. Mark, L.S., Pittenger, J.B., Hines, H., Carello, C., Shaw, R.E., Todd, J.T.: Wrinkling and head shape as coordinated sources of age-level information. Perception & Psychophysics 27, 117–124 (1980)
3. Yamaguchi, M.K., Kato, T., Akamatsu, S.: Relationship between physical traits and subjective impressions of the face - Age and sex information. IEICE Trans. J79-A(2), 279–287 (1996)
4. Kanno, T., Akiba, M., Teramachi, Y., Nagahashi, H., Agui, T.: Classification of age Group Based on Facial Images of Young Males by Using Neural Networks. IEICE Trans. Inf. & Syst. E84-D(8), 1094–1101 (2001)
5. Kwon, Y.H., Lobo, N.D.V.: Age classification from facial images. In: CVPR 1994, Seattle, US, June 1994, pp. 762–767 (1994)
6. Burt, D.M., Perrett, D.I.: Preception of age in adult daudasian male faces: computer graphic manipulation of shape and colour information. Perception 259(1355), 137–143 (1995)
7. Ueki, K., Hayashida, T., Kobayashi, T.: Subspace-based age-group classification using facial images under various lighting conditions. In: Proc. of IEEE Intl. Conf. on Automatic Face and Gesture Recognition, pp. 43–48 (2006)
8. Takimoto, H., Mitsukura, Y., Fukumi, M., Akamatsu, N.: A Robust Gender and Age Estimation under Varying Facial Pose. IEEJ Trans. 127(7), 1022–1029 (2007)
9. Fujiwara, T., Koshimizu, H.: Age and Gender Estimations by Modeling Statistical Relationship among Faces. In: Palade, V., Howlett, R.J., Jain, L. (eds.) KES 2003. LNCS, vol. 2774, pp. 870–876. Springer, Heidelberg (2003)
10. Nagata, N., Inokuchi, S.: Subjective Age Obtained from Facial Images -How Old We Feel Compared to Others. In: Palade, V., Howlett, R.J., Jain, L. (eds.) KES 2003. LNCS, vol. 2774, pp. 877–881. Springer, Heidelberg (2003)
11. George, P.A., Hole, G.J.: Factors influencing the accuracy of age estimates of unfamiliar faces. Perception 24(1), 1059–1073 (1995)
12. http://www.hoip.softpia.pref.gifu.jp
13. Kawato, S., Tetsutani, N.: Circle-Frequency Filter and its Application. In: Proc. Int. Workshop on Advanced Image Technology, pp. 217–222
14. Kawaguchi, T., Hikada, D., Rizon, M.: Detection of the eyes from human faces by hough transform and separability filter. In: Proc. of ICIP 2000, pp. 49–52 (2000)

# Improving EEG Analysis by Using Paraconsistent Artificial Neural Networks

Jair Minoro Abe[1,2], Helder F.S. Lopes[2], and Kazumi Nakamatsu[3]

[1] Graduate Program in Production Engineering, ICET - Paulista University
R. Dr. Bacelar, 1212, CEP 04026-002 São Paulo – SP – Brazil
[2] Institute For Advanced Studies – University of São Paulo, Brazil
jairabe@uol.com.br, helder@autobyte.com.br
[3] School of Human Science and Environment/H.S.E. – University of Hyogo – Japan
nakamatu@shse.u-hyogo.ac.jp

**Abstract.** In this paper we present a study of EEG by using the Paraconsistent Artificial Neural Network – PANN that can manipulate imprecise, contradictory and paracomplete data. Some improvements for EEG analysis are discussed. Experimental results concerning Alzheimer Disease made are also reported.

**Keywords:** paraconsistent logic, annotated logic, artificial neural network, EEG, biomedicine and informatics, pattern recognition.

## 1 Introduction

The electroencephalogram - EEG is a brain electric signal activity register, resultant of the space-time representation of synchronic postsynaptic potentials. The most probable is that the main generating sources of these electric fields are perpendicularly guided regarding to the cortical surface, as the cortical pyramidal neurons [4].

The graphic registration of the sign of EEG can be interpreted as voltage flotation with mixture of rhythms, being frequently sinusoidal, ranging 1 to 70 Hz. In the clinical-physiologic practice, such frequencies are grouped in frequency bands: delta (0,5 to 4 Hz), theta (4,1 to 8,0 Hz), alpha (8,1 to 12,5 Hz), and beta (> 13 Hz). During the relaxed awake, normal EEG in adults is predominantly composed by alpha band frequency, which is generated by interactions of the slum-cortical and thalamocortical systems [4], [6].

One of the problems in EEG analysis, as well as any other measurements devices are limited and subjected to the inherent imprecision of the several sources involved: equipment, movement of the patient, electric registers and individual variability of physician visual analysis. Such imprecision can often include conflicting information or paracomplete data. Although several interesting theories have been developed in order to overcome such limitations, e.g. Fuzzy set theory, Rough theory, non-monotonic reasoning, among others, they cannot manipulate inconsistencies and paracompleteness, at least directely. So, we need a new kind of logic to deal with uncertainty, inconsistent and paracomplete data [5], [7].

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 466–473, 2008.

In this paper we add some improvements to a new type of Artificial Neural Network - ANN, namely Paraconsistent Artificial Neural Network - PANN [7] based on Paraconsistent Annotated Evidential Logic Eτ [5], which is capable of manipulating imprecise, inconsistent and paracomplete data. Such improvements are discussed and we show how PANN can be efficient in recognizing EEG standards. To illustrate this we mention its ability in Alzheimer Disease - AD diagnosis reported in [10].

## 2 Methodology

The process of wave analysis by PANN consists previously of:

### 2.1 Data Capturing

The capturing of the data is obtained from usual ones (magnetic or manually) and converted in vectors (finite sequence of numbers) [11].

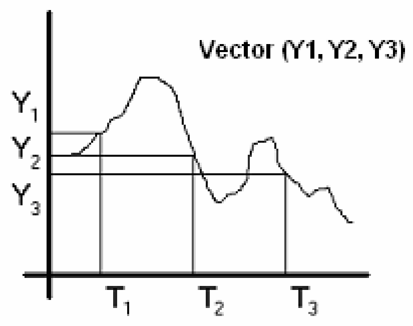


**Fig. 1.** Wave capturing through vectors

### 2.2 Adaptation for Screen Examination

As the actual EEG examination values can vary highly, in module, something 10 μV to 1500 μV, we proceed a normalization of the values between 100μV and -100 μV by a simple linear conversion, to facilitate the manipulation and to visualize in the screen:

$$x = \frac{100.a}{m} \tag{1}$$

where $m$ is the maximum value of the exam; $a$ is the current value of the exam. So, $x$ is the current normalized value.

### 2.3 Elimination of Negative Cycle

The minimum value of the exam is taken as zero value and the remaining values are translated proportionally.

### 2.4 Normalization for PANN Analysis

By a linear conversion, all data is normalized for PANN analysis.

It is worth to observe that the process above does not allow loss of any wave essential characteristics for our analysis.

## 3 Data Analysis, Expert System, and Wave Morphology

This expert system aims the analysis of the sign behavior in the morphologic aspect, that seeks to verify the format presented in the wave, where it is possible to verify different kinds of interference waveforms (artifacts) and spikes. This analysis also allows to verify the dominant frequency of the wave, verifying of which band it belongs (delta, theta, alpha and beta), by the control waves (normality pattern) they are stored with very defined frequencies.

The expert system will supply two output values: one favorable evidence (μ) and one contrary evidence (λ) according to the paraconsistent annotated evidential logic Eτ [].

With those two evidence values it is possible to obtain a resulting analysis by using the databases through logic Eτ structure.

In what follows, it is presented the characteristics of the analysis accomplished in this process.

### 3.1 Morphological Analysis

The process of the morphological analysis is accomplished comparing each point of the wave with all waves stored in the control database (waves with normal pattern). The wave that presents the maximum favorable evidence and the minimum contrary evidence will be chosen as the most similar wave to the wave that is being analyzed.

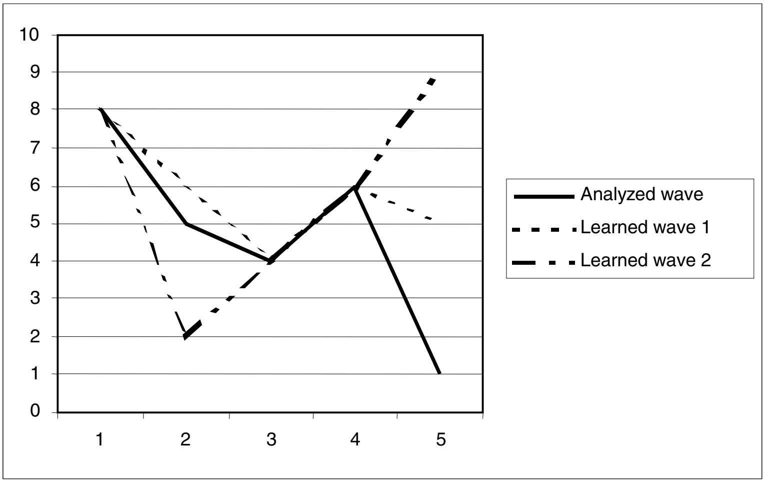


**Fig. 2.** Comparison among three waves

The control database is composed by waves presenting 256 positions with perfect sinusoidal morphology, with 0,5 Hz of variance, so taking into account Delta, Theta, Alpha and Beta (of 0,5 Hz to 30 Hz) wave groups.

In what follows, an example of the recognition process is presented using PANN:

Let's consider the following three waves.

**Table 1.** Table of the analyzed waves

| Wave's name | Position 1 | Position 2 | Position 3 | Position 4 | Position 5 |
|---|---|---|---|---|---|
| Analyzed wave | 8 | 5 | 4 | 6 | 1 |
| Learned wave 1 | 8 | 6 | 4 | 6 | 5 |
| Learned wave 2 | 8 | 2 | 4 | 6 | 9 |

The analyzed wave is the wave that will be submitted to PANN for recognition. The learned wave 1 and learned wave 2 were previously stored in the control database (normality pattern). Observe that 'visually', Learned wave 1 is more 'similar' than Learned wave 2. We introduce the concepts in order to deal mathematically this similarity.

The favorable evidence is obtained by the identical positions.

The contrary evidence is obtained by difference sum (in module) of the correspondent different positions.

Making a comparative between Analyzed wave and Learned wave 1, we have:

**Table 2.** Analyzed wave x Learned wave 1

| Wave's name | Position 1 | Position 2 | Position 3 | Position 4 | Position 5 | Total |
|---|---|---|---|---|---|---|
| Analyzed wave | 8 | 5 | 4 | 6 | 1 | - |
| Learned wave 1 | 8 | 6 | 4 | 6 | 5 | - |
| Favorable evidence | 1 | 0 | 1 | 1 | 0 | 3 |
| Contrary evidence | 0 | 1 | 0 | 0 | 4 | 5 |

Making a comparative between Analyzed wave and Learned wave 2, we have:

**Table 3.** Analyzed wave x Learned wave 2

| Wave's name | Position 1 | Position 2 | Position 3 | Position 4 | Position 5 | Total |
|---|---|---|---|---|---|---|
| Analyzed wave | 8 | 5 | 4 | 6 | 1 | - |
| Learned wave 2 | 8 | 2 | 4 | 6 | 9 | - |
| Favorable evidence | 1 | 0 | 1 | 1 | 0 | 3 |
| Contrary evidence | 0 | 3 | 0 | 0 | 8 | 11 |

Normalizing the values by the division of the favorable evidence ($\mu$) and of the contrary evidence ($\lambda$) for the number of elements of the wave, we have:

**Table 4.** Normalized values: Learned wave 1 and Learned wave2

| Case | μ | λ | Normalization FE | Normalization CE |
|---|---|---|---|---|
| Analyzed wave x Learned wave 1 | 3 | 5 | 0.6 | 1 |
| Analyzed wave x Learned wave 2 | 3 | 11 | 0.6 | 2.2 |

Therefore, we noticed that the wave with the maximum favorable evidence and the minimum contrary evidence is the learned Wave 1, in other words, this is the most similar wave to the analyzed Wave.

By this process, PANN was applied successfully in some studies, e.g. speech recognition [12]. However, in practice, we face with some new characteristics. That's the topic we concerned about next.

## 4 Counting the Number of Peaks

When the methodology is used in vectors with a huge number of positions, as it is the case of EEG analysis, it can present little variance among the differences found in the comparative study.

To avoid this, we introduce other characteristic factor of comparison, the number of peaks of the wave.

In this process, instead we consider as favorable evidence the equality between wave points, we substitute them for the proximity among the peaks of the analyzed waves:

$$1 - ((|bd - vt|) / (bd + vt)) \tag{2}$$

Where:

1. vt = number of wave peaks of the exam
2. bd = number of the wave peaks being compared (pattern stored in the database)

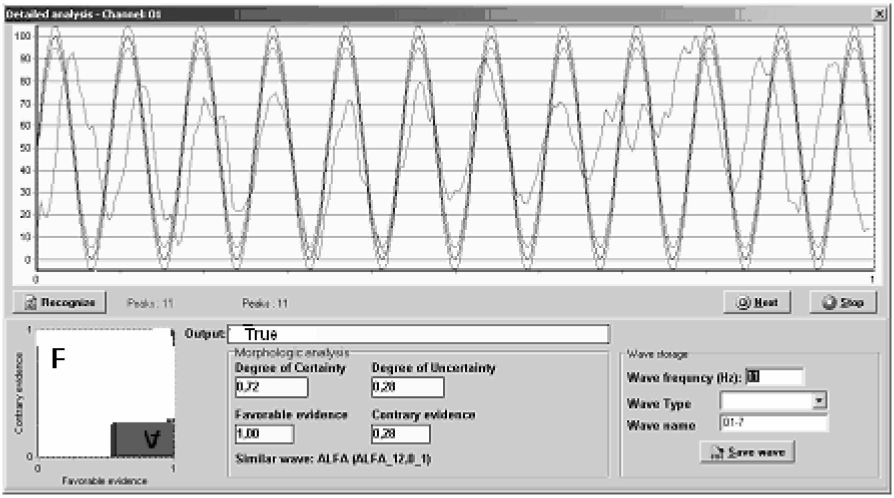
So, with this improvement we can detect 'difference' between waves more sharply allowing verifying different kinds of interference waveforms (artifacts) and spikes.

Another interesting information that can be obtained in this process it is the wave's approximate frequency. As the control waves of normality pattern were stored in the database in a systematic way, in other words, with waves with prefixed frequency. In this way, we know the frequency of each wave. Therefore, when we found the most similar wave to the wave that is being analyzed, we also found its frequency.

One most amazing advantage of this analysis method is the low processing, so it allows using relatively simpler mathematical techniques in comparison with the techniques used nowadays (such as FFT - Fast Fourier Transform).

### 4.1 Preliminary Tests

In what follows, we present an analysis by a software (Fig. 3). It shows an EEG exam (light grey) being compared with the most similar wave of the data group (slightly dark grey).



Output: True

Morphologic analysis

Degree of Certainty: 0,72

Degree of Uncertainty: 0,28

Favorable evidence: 1,00

Contrary evidence: 0,28

Similar wave: ALFA (ALFA_12,0_1)

Wave storage

Wave frequency (Hz): 0

Wave Type:

Wave name: 01-7

**Fig. 3.** Comparative between EEG wave and database waves (perfect sinusoidal waves)

**Table 5.** Lattice τ considered in the analysis

| Lattice τ | |
|---|---|
| V – True | $0.6 \leq \mu \leq 0.9$ & $\lambda \leq 0.28$ |
| | $\mu > 0.9$ & $\lambda < 0.28$ |
| | $0.62 \leq G_{ce} \leq 0.72$ & $0.28 < G_{ct} \leq 0.37$ |
| F – False | $\mu < 0.6$ & $\lambda > 0.37$; $G_{ce} \leq 0.62$ & $G_{ct} > 0.28$ |

The correspondent Cartesian representation is in Fig. 3 below.

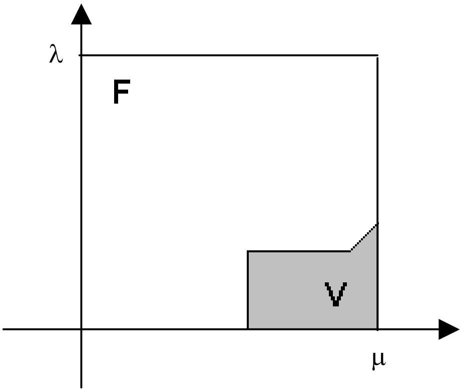


**Fig. 4.** Lattice τ used for the analysis

### 4.2 Experimental Results

As illustrative example, with suitable adaptations, an earlier version of the method was tested on real EEG exams in the aid of Alzheimer Disease diagnosis. Detailed exposition and discussion is to be found in [10]. Here we mention only the performance results. 10 EEGs were tested and the system classified correctly as normal at a rate of 80% and 20% as false-positive. More 10 EEGs was tested and the system classified correctly as AD at a rate of 80 % and 20% as false-negative, as in the table 6 below [10].

**Table 6.** Diagnosis – Normal x Probable AD patients [10]

| | Gold Standard | | |
|---|---|---|---|
| | AD Patient | Non AD Patient | Total |
| Negative | 2 | 2 | 4 |
| Positive | 8 | 8 | 16 |
| Total | 10 | 10 | 20 |
| | **Gold Standard** | | |
| | AD Patient | Non AD Patient | Total |
| Negative | 10% | 10% | 20% |
| Positive | 40% | 40% | 80% |
| Total | 50% | 50% | 100% |
| Sensitivity: | 0,8 | | |
| Specificity: | 0,8 | | |

## 5 Conclusions

The improvements discussed in this paper surely will be useful in signal analysis with EEG characteristics (in particular for improvements in the study of [10]); however for other applications extra improvements are necessary. But this is normal at actual moment of research stage of PANN. We hope to say more in forthcoming papers.

**Acknowledgments.** The authors are grateful to Dr. Renato Anghinah for providing EEG real data from Reference Center of Behavioral Disturbances and Dementia (CEREDIC) of Medicine School of University of São Paulo, Brazil for analysis (although this was already reported in [10]) as well as his supporting in EEG and AD informations for the paper.

## References

1. Duffy, F.H., Albert, M.S., McNulty, G., Garvey, A.J.: Age differences in brain electrical activity of healthy subjects. Ann. Neural. 1984 16, 430–438 (1984)
2. Nuwer, M.R., Comi, G., Emerson, R., Fuglsang-Frederiksen, J., Guéri, T.M., Hinrichs, H., Ikeda, A., Luccas, F.J.C., Rappelsberger, P.: IFCN standards for digital recording of clinical EEG. Electroencephalogr. Clin. Neurophysiol. 106, 259–261 (1998)

3. Nitrini, R., Caramelli, P., Bottino, C.M., Damasceno, B.P., Brucki, S.M., Anghinah, R.: Academia Brasileira de Neurologia. Diagnosis of Alzheimer's disease in Brazil: diagnostic criteria and auxiliary tests. Recommendations of the Scientific Department of Cognitive Neurology and Aging of the Brazilian Academy of Neurology. Arq. Neuropsiquiatr. 2005 63(3A), 713–719 (2005)
4. Montenegro, M.A., Cendes, F., Guerreiro, M.M., Guerreiro, C.A.M.: EEG na Prática Clínica. LEMOS Editorial 2001, 303 (2001)
5. Abe, J.M.: Fundamentos da Lógica Anotada (in Portuguese), Ph.D. Thesis, FFLCH - USP, São Paulo, Brazil, p. 135 (1992)
6. Anghinah, R.: Estudo da densidade espectral e da coerência do eletrencefalograma em indivíduos adultos normais e com doença de Alzheimer provável (in Portuguese), Ph.D. Thesis, Faculdade de Medicina da Universidade de São Paulo, São Paulo, Brazil (2003)
7. Da Silva Filho, J.I., Abe, J.M.: Fundamentos das Redes Neurais Paraconsistentes – Destacando Aplicações em Neurocomputação. Editora Arte & Ciência 2001, p. 247 (2001)
8. Fausett, L.: Fundamentals of Neural Network Architectures, Algorithms, and Applications, S. edn., USA. Prentice Hall, Englewood Cliffs (1994)
9. Baxt, W.J.: Application of Artificial Neural Network to Clinical Medicine. Lancet. 346, 1135–1138 (1995)
10. Abe, J.M., Lopes, H.F.S., Anghinah, R.: Paraconsistent Artificial Neural Network and Alzheimer Disease: A Preliminary Study. Dementia & Neuropsychologia 3, 241–247 (2007)
11. Abe, J.M., Prado, J.C.A., Nakamatsu, K.: Paraconsistent Artificial Neural Network: Applicability in Computer Analysis of Speech Productions. In: Gabrys, B., Howlett, R.J., Jain, L.C. (eds.) KES 2006. LNCS (LNAI), vol. 4252, pp. 844–850. Springer, Heidelberg (2006)
12. Prado, J.C.A.: Constelação Fônica e Redes Neurais Artificiais: Aplicabilidade na Análise Computacional da Produção da Fala. Ph.D. Thesis, University od São Paulo, São Paulo, Brazil (2007)

# Transitive Reasoning of Before-After Relation Based on Bf-EVALPSN

Kazumi Nakamatsu[1], Jair Minoro Abe[2], and Seiki Akama[3]

[1] University of Hyogo, Himeji, Japan
nakamatu@shse.u-hyogo.ac.jp
[2] Paulista University, Sao Paulo, Brazil
jairabe@uol.com.br
[3] C-republic, Kawasaki, Japan
sub-akama@jcom.home.ne.jp

**Abstract.** A paraconsistent annotated logic program called bf-EVALPSN has been developed for dealing with before-after relations between processes and applied to real-time process order control. In order to increase the efficiency of bf-EVALPSN process order control, a transitive reasoning system of before-after relations in bf-EVALPSN is introduced.

**Keywords:** process order control, EVALPSN, bf-EVALPSN, before-after relation, paraconsistent reasoning system.

## 1 Introduction

We have already developed a paraconsistent annotated logic program called Extended Vector Annotated Logic Program with Strong Negation(abbr. EVALPSN), which has been applied to various kinds of process safety verification and control such as pipeline process control [3,4,5]. We have also developed an EVALPSN called bf(before-after)-EVALPSN to deal with bf(before-after)-relations between time intervals, and applied it to real-time process order control. In bf-EVALPSN process order control, a particular EVALPSN literal $R(pi, pj, t):[(i, j), \mu]$ is used to represent the bf-relation between processes $pi$ and $pj$ at time $t$, and bf-relations are determined in real-time according to the order of process start/finish times. Suppose that we deal with $n$ processes in a bf-EVALPSN process order control system, then there are ${}_n\mathrm{C}_2$ bf-relations to be dealt with according to each start/finish information of processes. Since it is not so efficient to deal with all bf-relations, if we have a transitive reasoning system for bf-relations, which can reason all bf-relations from neighbor bf-relations in real-time, the performance of bf-EVALPSN process order control would be incresed. Exactly speaking of transitive reasoning, the bf-relation between processes $Pr_i$ and $Pr_k$ can be reasoned from two bf-relations between processes $Pr_i$ and $Pr_j$, and between processes $Pr_j$ and $Pr_k$ transitively, where $i < j < k$. In this paper, we introduce some bf-EVALPSN inference rules for the transitive reasoning system.

This paper is organized in the following manner: firstly, EVALPSN is reviewed briefly; next bf-EVALPSN and its implementation are introduced with a simple example; lastly, some bf-EVALPSN inference rules are introduced.

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 474–482, 2008.

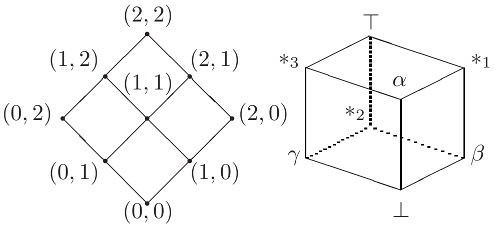


**Fig. 1.** Lattice $\mathcal{T}_v(2)$ and Lattice $\mathcal{T}_d$

## 2 EVALPSN

We review EVALPSN briefly[4]. Generally, a truth value called an *annotation* is explicitly attached to each literal in annotated logic programs [1]. For example, let $p$ be a literal, $\mu$ an annotation, then $p:\mu$ is called an *annotated literal.* The set of annotations constitutes a complete lattice. An annotation in EVALPSN has a form of $[(i,j),\mu]$ called an *extended vector annotation.* The first component $(i,j)$ is called a *vector annotation* and the set of vector annotations constitutes the complete lattice, $\mathcal{T}_v(n) = \{\ (x,y)|0 \le x \le n, 0 \le y \le n, x, y$ and $n$ are integers $\}$ in Fig.1. The ordering($\preceq_v$) of $\mathcal{T}_v(n)$ is defined as : let $(x_1,y_1)$, $(x_2,y_2) \in \mathcal{T}_v(n)$, $(x_1,y_1) \preceq_v (x_2,y_2)$ iff $x_1 \le x_2$ and $y_1 \le y_2$. For each extended vector annotated literal $p:[(i,j),\mu]$, the integer $i$ denotes the amount of positive information to support the literal $p$ and the integer $j$ denotes that of negative one. The second component $\mu$ is an index of fact and deontic notions such as obligation, and the set of the second components constitutes the complete lattice, $\mathcal{T}_d = \{\bot, \alpha, \beta, \gamma, *_1, *_2, *_3, \top\}$. The ordering($\preceq_d$) of $\mathcal{T}_d$ is described by the Hasse's diagram in Fig.1. The intuitive meaning of each member of $\mathcal{T}_d$ is $\bot$ (unknown), $\alpha$ (fact), $\beta$ (obligation), $\gamma$ (non-obligation), $*_1$ (fact and obligation), $*_2$ (obligation and non-obligation), $*_3$ (fact and non-obligation), and $\top$ (inconsistency). Then the complete lattice $\mathcal{T}_e(n)$ of extended vector annotations is defined as the product $\mathcal{T}_v(n) \times \mathcal{T}_d$. The ordering($\preceq_e$) of $\mathcal{T}_e(n)$ is defined as : let $[(i_1,j_1),\mu_1]$ and $[(i_2,j_2),\mu_2] \in \mathcal{T}_e$, $[(i_1,j_1),\mu_1] \preceq_e [(i_2,j_2),\mu_2]$ iff $(i_1,j_1) \preceq_v (i_2,j_2)$ and $\mu_1 \preceq_d \mu_2$.

There are two kinds of *epistemic negation* ($\neg_1$ and $\neg_2$) in EVALPSN, which are defined as mappings over $\mathcal{T}_v(n)$ and $\mathcal{T}_d$, respectively.

**Definition 1**(epistemic negations $\neg_1$ and $\neg_2$ in EVALPSN)
$\neg_1([(i,j),\mu]) = [(j,i),\mu],\ \forall\mu \in \mathcal{T}_d,\ \neg_2([(i,j),\bot]) = [(i,j),\bot]$,
$\neg_2([(i,j),\alpha]) = [(i,j),\alpha],\ \neg_2([(i,j),\beta]) = [(i,j),\gamma],\ \neg_2([(i,j),\gamma]) = [(i,j),\beta]$,
$\neg_2([(i,j),*_1]) = [(i,j),*_3],\ \neg_2([(i,j),*_2]) = [(i,j),*_2],\ \neg_2([(i,j),\top]) = [(i,j),\top]$.

If we regard the epistemic negations as syntactical operations, the epistemic negations followed by literals can be eliminated by the syntactical operations. For example, $\neg_1 p:[(2,0),\alpha] = p:[(0,2),\alpha]$ and $\neg_2 q:[(1,0),\beta] = p:[(1,0),\gamma]$.

There is another negation called *strong negation* ($\sim$) in EVALPSN, and it is treated as classical negation.

**Definition 2** (strong negation $\sim$) [2]**.** Let $F$ be any formula and $\neg$ be $\neg_1$ or $\neg_2$. $\sim F =_{def} F \rightarrow ((F \rightarrow F) \wedge \neg(F \rightarrow F))$.

**Definition 3** (well extended vector annotated literal)**.** Let $p$ be a literal. $p : [(i,0),\mu]$ and $p : [(0,j),\mu]$ are called *weva(well extended vector annotated)-literals*, where $i,j \in \{1,2,\cdots,n\}$, and $\mu \in \{\ \alpha,\ \beta,\ \gamma\ \}$.

**Definition 4** (EVALPSN)**.** If $L_0,\cdots,L_n$ are weva-literals, $L_1 \wedge \cdots \wedge L_i \wedge \sim L_{i+1} \wedge \cdots \wedge \sim L_n \rightarrow L_0$ is called an *EVALPSN clause*. An *EVALPSN* is a finite set of EVALPSN clauses.

Fact and deontic notions, “obligation”, “forbiddance” and “permission” are represented by extended vector annotations, $[(m,0),\alpha]$, $[(m,0),\beta]$, $[(0,m),\beta]$, and $[(0,m),\gamma]$, respectively, where $m$ is a positive integer.

## 3 Before-After EVALPSN(bf-EVALPSN)

First of all, we introduce a particular literal $R(pi,pj,t)$ whose vector annotation represents the before-after relation between processes $Pr_i(pi)$ and $Pr_j(pj)$ at time $t$, and it is called a *bf-literal* [1] .

**Definition 5** (bf-EVALPSN)**.** An extended vector annotated literal

$$R(p_i,p_j,t) : [(i,j),\mu]$$

is called a *bf-EVALP literal*, where $(i,j)$ is a vector annotation and $\mu \in \{\alpha,\beta,\gamma\}$. If an EVALPSN clause contains bf-EVALP literals, it is called a *bf-EVALPSN clause* or just a *bf-EVALP clause* if it contains no strong negation. A *bf-EVALPSN* is a finite set of bf-EVALPSN clauses.

We provide a paraconsistent before-after interpretation for vector annotations to represent bf-relations in bf-EVALPSN, and such vector annotations are called *bf-annotations*. Exactly speaking, bf-relations between processes are classified into meaningful fifteen kinds according to bf-relations between start/finish times of two processes in bf-EVALPSN though[6], we consider ten kinds among the fifteen ones for simplicity in this paper.

**Before (`be`)/After (`af`)**

First of all, we define the most basic bf-relations *before*/*after* according to the bf-relation between each start time of two processes, which are represented by bf-annotations `be`/`af`, respectively. Suppose that there are two processes, $Pr_i$ with start time $x_s$ and finish time $x_f$, and $Pr_j$ with start time $y_s$ and finish time $y_f$. If one process has started before/after another one starts, then the bf-relations between them are defined as “before(`be`)/after(`af`)”, respectively. They are described in the process time chart Figure 2 with the condition that process $Pr_i$ has started before process $Pr_j$ starts. The bf-relation between their

[1] Hereafter, the word “**b**efore-a**f**ter” is abbreviated as just “bf” in this paper.

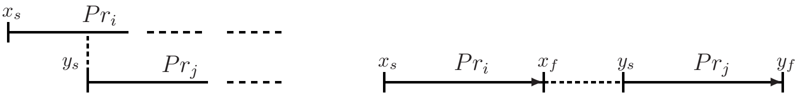


**Fig. 2.** Before/After and Disjoint Before/After

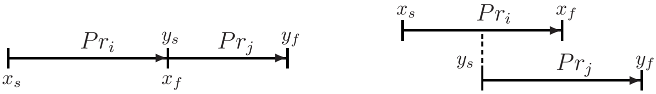


**Fig. 3.** Immediate Before/After and Joint Before/After

start/finish times is denoted by the inequality $\{x_s < y_s\}$ [2]. For example, a fact at time $t$ "process $Pr_i$ has started before process $Pr_j$ starts" can be represented by the bf-EVALP clause, $R(pi, pj, t) : [\texttt{be}, \alpha]$.

We define other eight kinds of bf-annotations as well as before(be)/after(af).

**Disjoint Before (db)/After (da)**
bf-relations *disjoint before*(db)/ *after*(da) are described in Fig.2.
**Immediate Before (mb)/After (ma)**
bf-relations *immediate before*(mb)/ *after*(ma) are described in Fig.3.
**Joint Before (jb)/After (ja)**
bf-relations *joint before*(jb)/ *after*(ja) are described in Fig.3.
**Included Before (ib)/After (ia)**
bf-relations *included before*(ib)/ *after*(ia) in Fig.4.

If we take before-after measure over the ten bf-annotations as the horizontal order and before-after knowledge amount of them as the vertical one, we obtain the complete bi-lattice $\mathcal{T}_v(7)_{bf}$ of bf-annotations in Fig.4. Then, there is the following correspondence between bf-annotations and vector annotations:

$$\texttt{be}(0,4)/\texttt{af}(4,0), \quad \texttt{db}(0,7)/\texttt{da}(7,0), \quad \texttt{mb}(1,6)/\texttt{ma}(6,1),$$
$$\texttt{jb}(2,5)/\texttt{ja}(5,2), \quad \texttt{ib}(3,4)/\texttt{ia}(4,3).$$

**Definition 6** ($\neg_1$ in bf-EVALPSN)**.** Obviously epistemic negation $\neg_1$ that maps bf-annotations { be, af, da, db, ma, mb, ja, jb, ia, ib } to themselves is defined as follows:

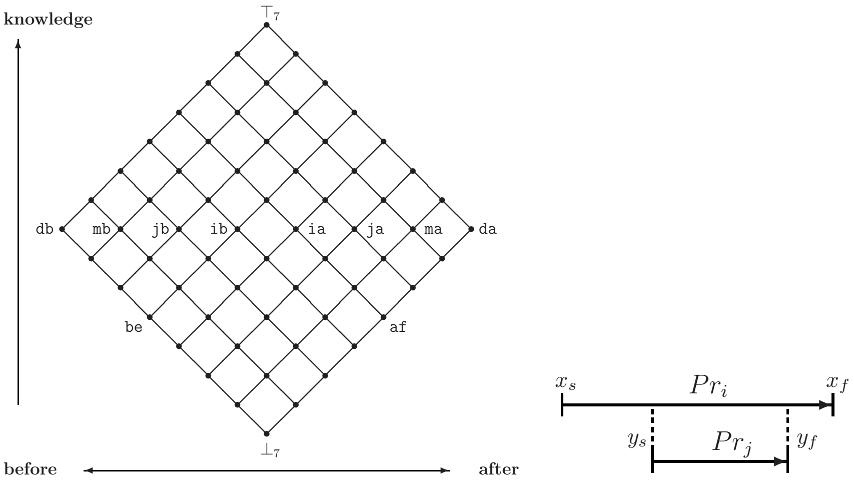
$$\neg_1(\texttt{af}/\texttt{be}) = \texttt{be}/\texttt{af}, \quad \neg_1(\texttt{da}/\texttt{db}) = \texttt{db}/\texttt{da}, \quad \neg_1(\texttt{ma}/\texttt{mb}) = \texttt{mb}/\texttt{ma},$$
$$\neg_1(\texttt{ja}/\texttt{jb}) = \texttt{jb}/\texttt{ja}, \quad \neg_1(\texttt{ia}/\texttt{ib}) = \texttt{ib}/\texttt{ia}, \quad \neg_1(\bot_7/\top_7) = \bot_7/\top_7.$$

## 4 Before-After Relation Computing

We consider two bf-relations(bf-EVALPSN clauses),

$$R(Pr_0, Pr_1, t) : [(i_1, j_1), \alpha] \quad \text{and} \quad R(Pr_1, Pr_2, t) : [(i_2, j_2), \alpha]$$

[2] If time $t_1$ is earlier than time $t_2$, we conveniently denote the relation by the inequality $t_1 < t_2$.



Lattice $\mathcal{T}_v(7)_{bf}$, Included Before/After

**Fig. 4.**

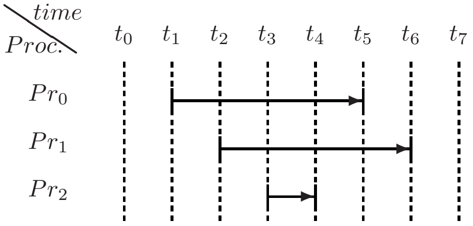


**Fig. 5.** Process Schedule

between three processes $Pr_0$, $Pr_1$ and $Pr_2$ scheduled in Fig.5, and describe how the bf-relations are computed according to process start/finish time sequence $t_0, \cdots, t_7$.

**At time** $t_0$, no process has started, thus we have no knowledge about both the bf-relations. Therefore, we have the bf-EVALPSN clauses,

$$R(Pr_0, Pr_1, t_0):[(0,0), \alpha], \quad \text{and} \quad R(Pr_1, Pr_2, t_0):[(0,0), \alpha].$$

**At time** $t_1$, only process $Pr_0$ has started, then it can be reasoned that the bf-relation between processes $Pr_0$ and $Pr_1$ is one of bf-relations $\{\texttt{db}(0,7), \texttt{mb}(1,6), \texttt{jb}(2,5), \texttt{ib}(3,4)\}$ whose greatest lower bound is $(0,4)$. On the other hand, we have no knowledge about the bf-relation between processes $Pr_1$ and $Pr_2$. Therefore, we have the bf-EVALPSN clauses,

$$R(Pr_0, Pr_1, t_1):[(0,4), \alpha], \quad \text{and} \quad R(Pr_1, Pr_2, t_1):[(0,0), \alpha].$$

**At time** $t_2$, process $Pr_1$ has started before process $Pr_0$ finishes, then it can be reasoned that the bf-relation between processes $Pr_0$ and $Pr_1$ is one of bf-relations

$\{\texttt{jb}(2,5), \texttt{ib}(3,4)\}$ whose greatest lower bound is $(2,4)$. Moreover, as process $Pr_2$ has not started yet literal $R(Pr_1, Pr_2, t_2)$ has the vector annotation $(0,4)$ as well as literal $R(Pr_0, Pr_1, t_1)$. Therefore, we have the bf-EVALPSN clauses,

$$R(Pr_0, Pr_1, t_2):[(2,4), \alpha], \quad \text{and} \quad R(Pr_1, Pr_2, t_2):[(0,4), \alpha].$$

**At time** $t_3$, process $Pr_2$ has started before processes $Pr_0$ and $Pr_1$ finish, then it can be reasoned that literals $R(Pr_0, Pr_1, t_3)$ and $R(Pr_2, Pr_3, t_3)$ have the same vector annotation $(2,4)$ as well as literal $R(Pr_0, Pr_1, t_2)$. Therefore, we have the bf-EVALPSN,

$$R(Pr_0, Pr_1, t_3):[(0,0), \alpha], \quad \text{and} \quad R(Pr_1, Pr_2, t_3):[(0,0), \alpha].$$

**At time** $t_4$, process $Pr_2$ has finished before processes $Pr_1$ and $Pr_2$ finish, then literal $R(Pr_0, Pr_1, t_4)$ still has the same vector annotation $(2,4)$ at time $t_3$ because neither processes $Pr_0$ nor $Pr_1$ has finished yet, and literal $R(Pr_0, Pr_1, t_4)$ has its final vector annotation $\texttt{ib}(3,4)$. Therefore, we have the bf-EVALPSN clause,

$$R(Pr_0, Pr_1, t_4):[(3,4), \alpha], \quad \text{and} \quad R(Pr_1, Pr_2, t_4):[(2,4), \alpha].$$

**At time** $t_5$, process $Pr_0$ has finished before processes $Pr_1$ finish, then literal $R(Pr_1, Pr_2, t_5)$ has its final vector annotation $\texttt{jb}(2,5)$. Therefore, we have confirmed bf-relations(bf-EVALPSN clause),

$$R(Pr_0, Pr_1, t_5):[(3,4), \alpha], \quad \text{and} \quad R(Pr_1, Pr_2, t_5):[(2,5), \alpha].$$

## 5 Transitive Reasoning Based on Bf-EVALPSN

In this section, we consider three processes $Pr_0$, $Pr_1$ and $Pr_2$, and three kinds of bf-relations between those processes in order to derive bf-EVALPSN inference rules that can logically reason bf-EVALPSN clause $R(p0, p2, t):[(i_2, j_2), \alpha]$ from bf-EVALPSN clauses $R(p0, p1, t):[(i_0, j_0), \alpha]$ and $R(p1, p2, t):[(i_2, j_2), \alpha]$ in three process time charts 1,2,3 Fig.6. In those process time charts only the start time of process $Pr_2$ is varying between times $t_3$ and $t_5$, and three kinds of variation of vector annotations representing bf-relations are shown in Table 1. For each

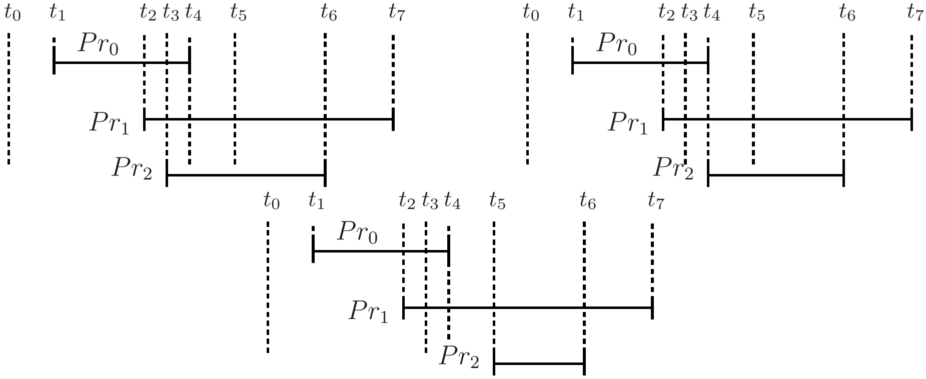


**Fig. 6.** Process Time Chart 1(top left), 2(top right), 3(bottom left)

**Table 1.** Vector Annotations of Process Time Chart 1,2,3

| process time chart 1 | $t_0$ | $t_1$ | $t_2$ | $t_3$ | $t_4$ | $t_5$ | $t_6$ | $t_7$ |
|---|---|---|---|---|---|---|---|---|
| $R(p0, p1, t)$ | $(0, 0)$ | $(0, 4)$ | $(2, 4)$ | $(2, 4)$ | $(2, 5)$ | $(2, 5)$ | $(2, 5)$ | $(2, 5)$ |
| $R(p1, p2, t)$ | $(0, 0)$ | $(0, 0)$ | $(0, 4)$ | $(2, 4)$ | $(2, 4)$ | $(2, 4)$ | $(3, 4)$ | $(3, 4)$ |
| $R(p0, p2, t)$ | $(0, 0)$ | $(0, 4)$ | $(0, 4)$ | $(2, 4)$ | $(2, 5)$ | $(2, 5)$ | $(2, 5)$ | $(2, 5)$ |
| process time chart 2 | $t_0$ | $t_1$ | $t_2$ | $t_3$ | $t_4$ | $t_5$ | $t_6$ | $t_7$ |
| $R(p0, p1, t)$ | $(0, 0)$ | $(0, 4)$ | $(2, 4)$ | $(2, 4)$ | $(2, 5)$ | $(2, 5)$ | $(2, 5)$ | $(2, 5)$ |
| $R(p1, p2, t)$ | $(0, 0)$ | $(0, 0)$ | $(0, 4)$ | $(0, 4)$ | $(2, 4)$ | $(2, 4)$ | $(3, 4)$ | $(3, 4)$ |
| $R(p0, p2, t)$ | $(0, 0)$ | $(0, 4)$ | $(0, 4)$ | $(0, 4)$ | $(1, 6)$ | $(1, 6)$ | $(1, 6)$ | $(1, 6)$ |
| process time chart 3 | $t_0$ | $t_1$ | $t_2$ | $t_3$ | $t_4$ | $t_5$ | $t_6$ | $t_7$ |
| $R(p0, p1, t)$ | $(0, 0)$ | $(0, 4)$ | $(2, 4)$ | $(2, 4)$ | $(2, 5)$ | $(2, 5)$ | $(2, 5)$ | $(2, 5)$ |
| $R(p1, p2, t)$ | $(0, 0)$ | $(0, 0)$ | $(0, 4)$ | $(0, 4)$ | $(0, 4)$ | $(2, 4)$ | $(3, 4)$ | $(3, 4)$ |
| $R(p0, p2, t)$ | $(0, 0)$ | $(0, 4)$ | $(0, 4)$ | $(0, 4)$ | $(0, 7)$ | $(0, 7)$ | $(0, 7)$ | $(0, 7)$ |

table, if we focus on the vector annotations at times $t_1$ and $t_2$, the following bf-EVALPSN inference rule can be derived:

**rule-1**

$$R(p0, p1, t) : [(0, 4), \alpha] \wedge R(p1, p2, t) : [(0, 0), \alpha] \rightarrow R(p0, p2, t) : [(0, 4), \alpha], \quad (1)$$

Furthermore, if we also focus on the vector annotations at times $t_3$ and $t_4$ in Table 1, the following two bf-EVALPSN rules also can be derived:

**rule-2**

$$R(p0, p1, t) : [(2, 4), \alpha] \wedge R(p1, p2, t) : [(2, 4), \alpha] \rightarrow R(p0, p2, t) : [(2, 4), \alpha], \quad (2)$$

**rule-3**

$$R(p0, p1, t) : [(2, 5), \alpha] \wedge R(p1, p2, t) : [(2, 4), \alpha] \rightarrow R(p0, p2, t) : [(2, 5), \alpha]. \quad (3)$$

As well as **rule-2** and **rule-3**, the following two rules also can be derived with focusing on the variations of vector annotations at time $t_4$ in Table 5.

**rule-4**

$$R(p0, p1, t) : [(2, 5), \alpha] \wedge R(p1, p2, t) : [(2, 4), \alpha] \rightarrow R(p0, p2, t) : [(1, 6), \alpha], \quad (4)$$

**rule-5**

$$R(p0, p1, t) : [(2, 5), \alpha] \wedge R(p1, p2, t) : [(0, 4), \alpha] \rightarrow R(p0, p2, t) : [(0, 7), \alpha]. \quad (5)$$

Here let us take **rule-3** and **rule-4**, since they have the same precedent,

$$R(p0, p1, t) : [(2, 5), \alpha] \wedge R(p1, p2, t) : [(2, 4), \alpha],$$

and different consequents,

$$R(p0, p2, t) : [(2, 5), \alpha] \quad \text{and} \quad R(p0, p2, t) : [(1, 6), \alpha],$$

those inference rules cannot be uniquely applied. Therefore, we need to consider applicable orders of bf-EVALPSN inference rules. Obviously there are three applicable orders (6), (7) and (8) of the rules.

$$\textbf{Order-1} \qquad \textbf{rule-1} \longrightarrow \textbf{rule-2} \longrightarrow \textbf{rule-3} \tag{6}$$

$$\textbf{Order-2} \qquad \textbf{rule-1} \longrightarrow \textbf{rule-4} \tag{7}$$

$$\textbf{Order-3} \qquad \textbf{rule-1} \longrightarrow \textbf{rule-5} \tag{8}$$

Let us show an application of bf-EVALPSN inference rules with taking the process time chart 3 in Fig.6 as an example.

**At time** $t_1$, **rule-1** is applied and we have the bf-EVALPSN clause,

$$R(p0, p2, t_1) : [(0, 4), \alpha].$$

**At times** $t_2$ and $t_3$, no rule can be applied and we still have the same vector annotation as

$$R(p0, p2, t_3) : [(0, 4), \alpha].$$

**At time** $t_4$, only **rule-5** can be applied and we obtain the bf-EVALPSN clause,

$$R(p0, p2, t_4) : [(0, 7), \alpha]$$

and the bf-relation between processes $Pr_0$ and $Pr_2$ has been resoned by bf-EVALPSN inference rules ordered in (8).

Due to space restriction we could not introduce all bf-EVALPSN inference rules and the proofs of their soundness and completeness though, we may save the computing cost of bf-EVALPSN process order control if bf-EVALPSN inference rules are used.

## 6 Conclusions

In this paper, we have introduced an annotated logic program called bf-EVALPSN, its implementation for process order control with a simple example, and bf-EVALPSN inference rules that can deal with transitive reasoning of bf-relations in EVALPSN.

## References

1. Blair, H.A., Subrahmanian, V.S.: Paraconsistent Logic Programming. Theoretical Computer Science 68, 135–154 (1989)
2. da Costa, N.C.A., Subrahmanian, V.S., Vago, C.: The Paraconsistent Logics P$\tau$. Zeitschrift für Mathematische Logic und Grundlangen der Mathematik 37, 139–148 (1989)
3. Nakamatsu, K.: Pipeline Valve Control Based on EVALPSN Safety Verification. J. Advanced Computational Intelligence and Intelligent Informatics 10, 647–656 (2006)

4. Nakamatsu, K., Abe, J.M., Suzuki, A.: Annotated Semantics for Defeasible Deontic Reasoning. In: Ziarko, W., Yao, Y. (eds.) RSCTC 2000. LNCS (LNAI), vol. 2005, pp. 432–440. Springer, Heidelberg (2001)
5. Nakamatsu, K., Mita, Y., Shibata, T.: An Intelligent Action Control System Based on Extended Vector Annotated Logic Program and its Hardware Implementation. J. Intelligent Automation and Soft Computing 13, 289–304 (2007)
6. Nakamatsu, K., Abe, J.M., Akama, S.: Paraconsistent Before-after Relation Reasoning Based on EVALPSN. In: The First Int'l. Symposium on Intelligent Interactive Multimedia Systems and Services. LNCS (LNAI). Springer, Heidelberg (to appear, 2008)

# A Conceptual Model for Guiding the Clustering Analysis

Wagner F. Castilho[1,4], Gentil J. Lucena Filho[2], Hércules A. do Prado[2,3], Edilson Ferneda[2], and Margarete Axt[4]

[1] Brazilian Federal Savings Bank, Brasília, DF – Brazil
SRTVN 701, conjunto C, Bloco A – Sala 321
70.719-930 Brasília, DF – Brazil
[2] Graduate Program in Knowledge and Information Technology Management
Catholic University of Brasília (UCB)
SGAN 916, Módulo B
91.501-970 Brasília, DF – Brazil
[3] Embrapa Food Technology – CTAA
Av. das Américas, 29501 - Guaratiba.
23.020-470 Rio de Janeiro, RJ – Brazil
[4] Federal University of Rio Grande do Sul
Av. Paulo Gama, 110
90.040-060 Porto Alegre, RS – Brazil
castilhowagner@gmail.com, glucena@pos.ucb.br,
hercules@ctaa.embrapa.br,
eferneda@pos.ucb.br, maaxt2002@ufrgs.br

**Abstract.** Knowledge discovery from databases, in the descriptive approach, includes clustering analysis (CA) as an alternative to estimate how a set of objects is organized in the space of their dimensions. The main objective in this task is to find "natural" groups that could exhibit some meaning. Considering the strong subjectivity that underlies this process, an important issue refers to the relationships among the CA players when looking for a model that could adjust the data. In this work, a model for actions coordination that provides an order to drive the relationships among CA players is presented. This model is presented as a conceptual contribution towards the construction of a computational environment to support effective conversations in a subjective context.

**Keywords:** Knowledge Discovery in Databases, Data mining, Clustering analysis, Action coordination.

## 1 Introduction

Departing from a set of objects, Clustering Analysis (CA) looks for a category structure that can fit in this data set. The aiming is to find "natural" groups, based in arbitrary internal criteria, in such a way that the cohesion among the members of a group would be the maximum and among the groups would be the minimum.

Grossly, the process of CA includes two basic steps: generating a clusters configuration and interpreting them in order to find some meaning in them. The first step is

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 483–490, 2008.

carried out by means of an algorithm, usually based in some kind of distance, which generates clouds of points. In the second step, specialists analyze these clouds aiming to find some meaning in the clusters. The second step presents a strong subjective bias, since it depends on mental models of the people (human beings) involved.

In this work we propose a model to deal with these subjective aspects in which a protocol based on speech acts is applied. This model provides a decision support process to build consensus and better articulated actions on the issues related to clusters interpretation.

The judgements and decisions from people involved with the process and the way they communicate on the elaboration of these thoughts and coordinate to make decisions, take actions and procedures is crucial for the planning cycle, execution and evaluation of the results from CA. These aspects can also be considered for application of data mining, multivariatre analysis, among others, guiding the relation between the people involved on the process.

## 2 An Overview on the Clustering Analysis

The whole CA process can be organized in nine steps (see Fig. 1): *(i)* domain and data understanding, *(ii)* definition of objectives, *(iii)* selection of relevant and discriminant variables, *(iv)* data preparation, *(v)* weighting definition, *(vi)* algorithm choice and configuration, *(vii)* algorithm application, *(viii)* results evaluation, and *(ix)* knowledge building and refining data structures. Notice that we assumed to apply a weighted clustering algorithm, as defined in [1].

In the first step a shared space of understanding about the domain and the data structure is built to enable the communication between the domain specialist and the data analyst. The former is related to the specific field in which the CA is being applied and the latter is the responsible for managing the whole CA process. While the domain specialist holds the knowledge regarding to the application area, the analyst master the methods, techniques and tools for CA. In the ideal situation they develop a synergy aiming to find a model that better adjust to the data.

In the second step, departing from a shared understanding space, they are guided to focus on defining the analysis objective.

In the third step the selection of variables are carried out taking into account their relevancy and how discriminant they are according to the analysis objective. Techniques like principal components analysis or factorial analysis [2] can be applied to figure out how discriminant is the selected variables. For short, low discriminant variables are those which values change very slightly among the objects, having a small effect in the clusters definition.

The fourth step is focussed in sampling, cleaning, and structuring the data set. The adequate treatment of missing values is also part of this step.

In the fifth step the components for the algorithm weighting is defined. In the informed clustering algorithm [1] an information matrix expressing the previous knowledge regarding to the application context and the data must be supplied as a way to introduce a domain bias in the clustering algorithm. This information matrix is built from a relationship (or cause-effect) mapping of the involved variables.

In the sixth step the clustering algorithm is chosen, according to the analyst or domain specialist negotiated preferences. The algorithm must be prepared to receive the information matrix, since it will provide the homogeneity coefficient that has to be considered in the clusters´ definition.

In the seventh step, the selected algorithm is applied in order to find a clustering configuration that can be seen as a candidate to represent the data structure. Many configurations can be generated until the specialist accepts it, according his experience in the domain.

In the eighth step the clustering results are evaluated. According to Cormack [3], many techniques exists that can be used to evaluate the quality of the generated clusters. There are two kinds of evaluation techniques for CA: the quantitative and

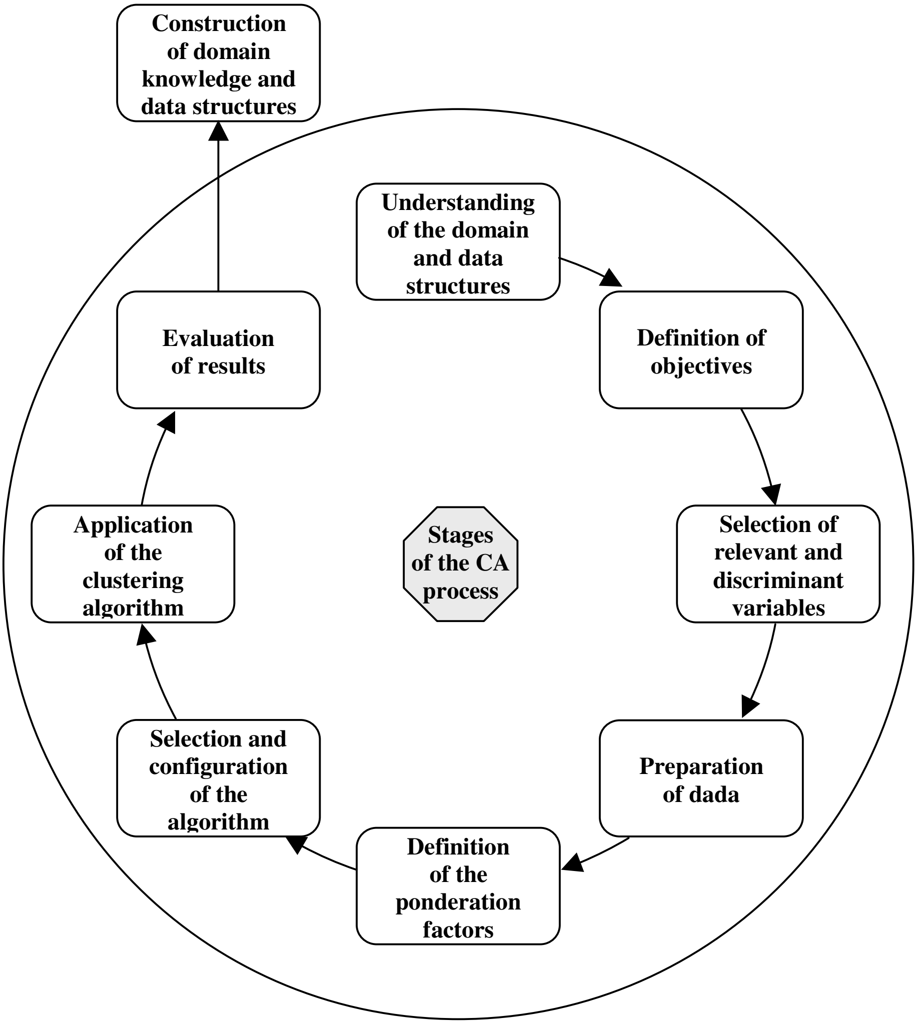

Construction of domain knowledge and data structures
Understanding of the domain and data structures
Definition of objectives
Selection of relevant and discriminant variables
Preparation of dada
Definition of the ponderation factors
Selection and configuration of the algorithm
Application of the clustering algorithm
Evaluation of results
Stages of the CA process

**Fig. 1.** Knowledge creation in clustering analysis

qualitative ones. As examples of quantitative techniques, Moreira [4], suggests the discriminant and the variance analysis. On the other hand, although, less precise, the qualitative approach cannot be ignored, since, by considering the huge amount of possible clustering configurations, one could argue that, in essence, the nature of the interpretation process is more qualitative than quantitative. According to this, in our view, the evaluation of results carried through this eighth step should consider both, the qualitative and quantitative approaches for this task.

The core of this paper is a roadmap to apply the qualitative approach that involves an intense and elaborated conversational agreement among the players. In the ninth step comprises the construction the knowledge that can include, beyond the application domain, the refinement of the own data structures. As it can be seen, this step is out the main cycle in Fig. 1. In a sense, this step can start another discovering cycle providing the input for the first step, in a spiral fashion.

## 3 The Actions Coordination Cycle

The conceptual basis for our proposal comes from [3], [4], [5], and [6], and is known as the actions coordination cycle. The actions coordination cycle has two phases: establishing a promise and promise accomplishment. The first one refers to the context creation and negotiation tasks, while the second one has to do with accomplishing the promise and the evaluation of the results derived from this accomplishment. There exist in the actions coordination cycle two agents involved when a promise situation occurs: the provider and the client.

The promise comprises the defined goals for the CA process. Precision and a explicit declaration for the customer is fundamental. Based on these defined (by the "client") and accepted (by the "service provider") goals, the results to be delivered should be marked with a statement of fulfillment in the form of a CA service accomplishment declaration. The client, once notified of this accomplishment declaration, should, in turn, declare a statement of satisfaction or dissatisfaction with the results just delivered, in accordance with his expectations presented at the begining.

An actions coordination cycle can be of two types, according to the nature of the speech act that starts it. It can be started by a request or by an offer. In both situations the provider and the client share a common space of interests and mutual commitments that is built from the expectations regarding the benefits that can come from the whole cycle. These expectations are supported by the reciprocal confidence that must permeate the relationship among the players.

Figs. 2 and 3 exhibit the schemas for the request and the offer cycles. In both cases a problem statement starts the cycle, beginning a context creation phase. In case of the request cycle, the problem statement is done by the client, based on his requirements for which satisfaction s/he depends on the provider. In case of the offer, the provider tries to meet what s/he figures out to be the client requirements.

Next, the negotiation phase starts after the request or offer statements have been posted and finishes with an acceptation statement. The acceptation statement in the request cycle is made by the provider and in the offer cycle is made by the client.

The next phase is the accomplishment, which begins with the promise statement and finishes with the accomplishment statement, always done by the provider. The

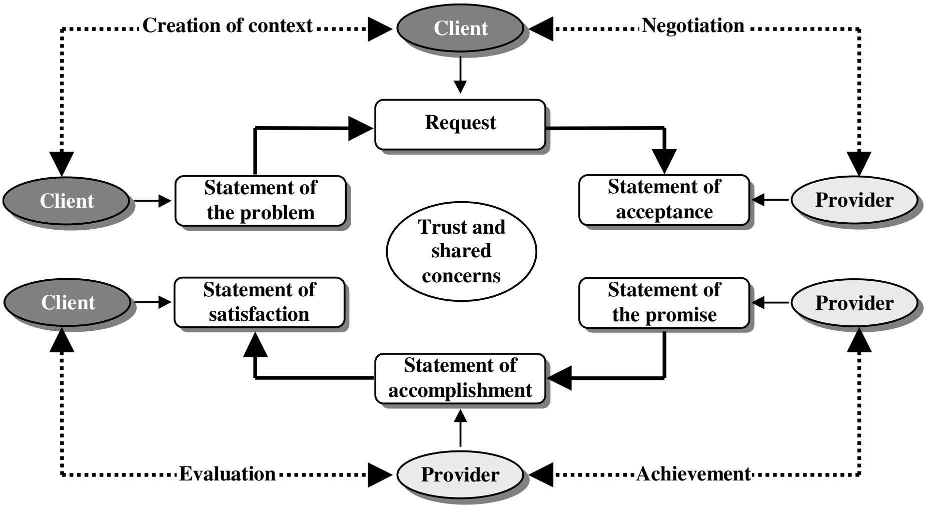


**Fig. 2.** The request cycle

fourth phase refers to the evaluation task and takes place after the provider declare the promise accomplishment, finishing with the satisfaction statement, always done by the client. This phase closes the request or offer cycles. However, not always these cycles end with the satisfaction statement. It may occur, depending how the previous phases were performed, that a client dissatisfaction statement may be expressed, closing those cycles in a non-effective way.

Notice that the differences between the request and the offer cycles are located in the upper side of the schemas. In the left-upper side of Fig. 2, the client behavior is characterized by thoughts regarding his necessities. Similarly, in Fig. 3, the provider is involved in thoughts related to the clients' necessities.

In the request cycle the client is in the two extremes of the context creation phase. He is responsible for the problem statement and for the sequence of speech acts (a conversation) that leads to the request. On the offer cycle, the provides plays a similar role, being in the two extremes of the context creation phase, when declaring the problem and the speech act that leads to the offer. These are the only important differences between the request and the offer cycles. In the lower sides of Figs. 2 and 3, the players' places and the nature of speech acts are the same.

The negotiation and evaluation phases are characterized by a bipolarity between the client and the provider, that are involved in a judgment sharing process in which an agreement with respect to the request or the offer is searched. Also, in this phase, a consensual evaluation of the promise accomplishment is desirable. These phases require parameters like action to be carried out, satisfaction conditions, and a timetable to accomplishment.

The context creation and the promise accomplishment phases are characterized by having only one player in their beginning and ending. For the request cycle, the context creation phase has the client in its both extremes and for the offer cycle this phase has the provider in its extremes. In addition, both cycles have the provider in the two extremes of the promise accomplishment phase.

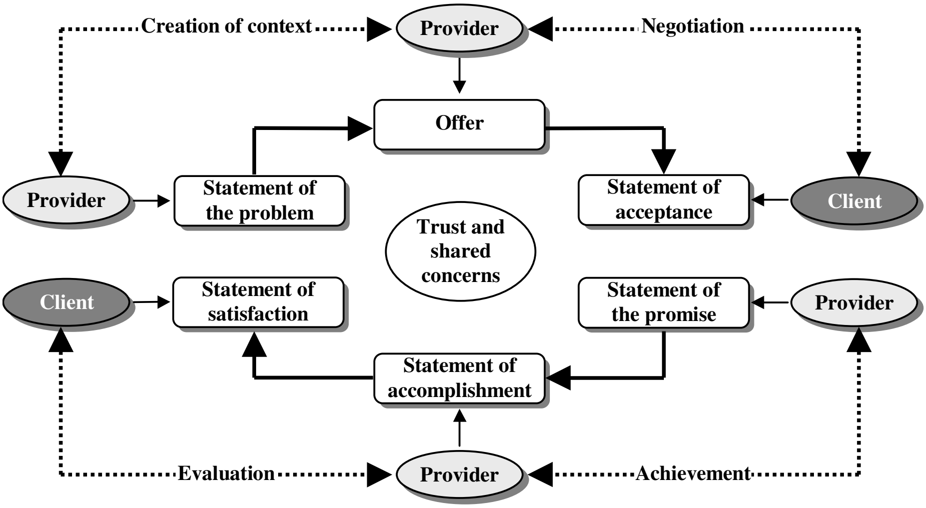


**Fig. 3.** The offer cycle

Notice that, in each phase of the actions coordination cycles it can be necessary to trigger new cycles in a commitment network, issuing, for example, new requests to other providers. This behavior was illustrated in Figs. 1 and 3 as entwined circles. In the heart of the cycles remains the shared confidence and concerns that are the basis for keeping the process cohesion. The weakening of these mutual feelings tends to provoke the process fragmentation.

## 4 Applying the Actions Coordination Cycle in CA

To approach the subjectivity in the CA process we propose to view it as an actions coordination cycle among the agents involved. The subjectivity in CA is mainly observed in the eighth and ninth steps of the process (results evaluation and knowledge building and refining data structures), since it is in those steps that human interpretations are more strongly present. However, it is important observe that, even in the other steps, there are different levels of subjectivity.

Ultimately speaking, the CA process, as any other process involving people, is a human process, that is, the subjectivity issue is not a peripheral one; it is central. So, we modeled the whole process applying the concepts presented in the previous section. An adapted schema from the actions coordination cycle to the CA process is shown in Fig. 4. It corresponds to the offer cycle in which the analyst plays the provider, while the domain specialist takes the place of a client. The analyst provides the knowledge creation from CA service.

The context creation phase corresponds to the domain and data structures understanding as a set up from the analyst to achieve a good interaction with the domain specialist. This interaction enables the next phase, the objectives definition. The analyst makes a first offer based in the necessities from the domain specialist and on the

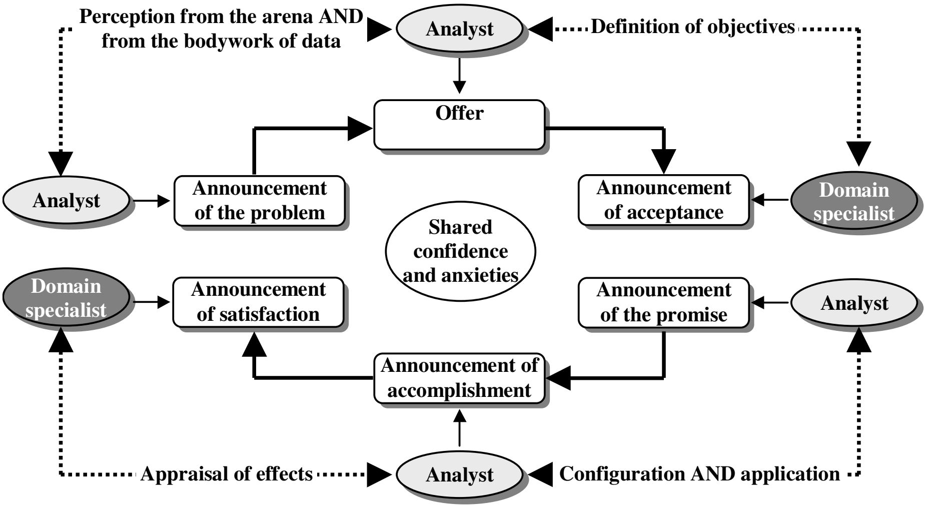

Perception from the arena AND from the bodywork of data
Analyst
Definition of objectives
Offer
Analyst
Announcement of the problem
Announcement of acceptance
Domain specialist
Shared confidence and anxieties
Domain specialist
Announcement of satisfaction
Announcement of the promise
Analyst
Announcement of accomplishment
Appraisal of effects
Analyst
Configuration AND application

**Fig. 4.** The actions coordination cycle applied to CA

knowledge acquired regarding to the problem context. This phase begins with the problem statement to the analyst and ends with the first offer he does.

In the objectives definition phase a shared space of knowledge is created between the analyst and the domain specialist. This phase corresponds to the negotiation phase in which the negotiation focus is the objectives to be seek during the CA process. It ends after an interaction between both players in order to meet an agreement that leads to the acceptance statement from the domain specialist.

In the configuration and application phase, which corresponds to the accomplishment phase in the offer cycle, the analyst performs the variables selection, the data preparation, the definition of the weighting factors, the choice of the algorithm and its configuration, as well its execution. This phase requires a strong interaction between the analyst and the domain specialist and is completed with the results presentation to evaluation, after a promise accomplishment statement from the analyst.

In the results evaluation phase the analyst and the domain specialist put their knowledge, judgments, and beliefs in action looking for an enlargement of the shared knowledge.

The actions coordination cycle in CA problem can be repeated many times, by redefining objectives, renegotiating agreements, and so on, until a satisfaction statement is obtained from the domain specialist.

## 5 Conclusions and Ongoing Work

According to Echeverría [5] when we talk about coordinating common actions, we are talking about communication. Among humans language is a recursive coordination of behavior based on reflection and reasoning. The same author states that “conversations are the effective component of linguistic interactions – the basic language units” and emphasizes the importance of the actions coordination in a world in which the auto-sufficiency is impossible. In this world, says Echeverría, we have to learn how to cooperate to coordinate actions. In this sense and in our point of view, the study and application of the actions coordination cycle in the CA process may help to promote a consensual understanding in a subjective learning context, enabling to feed a vast commitments network. The ongoing work includes both the application of this model for performance evaluation in public sanity companies and the development of an environment for conversation support in clustering analysis.

## References

1. Castilho, W.F., Prado, H.A., Ladeira, M.: Informed k-Means: a Clustering Process Biased by Prior Knowledge. In: Seruca, I., Filipe, J., Hammoudi, S., Cordeiro, J. (eds.) ICEIS: Proceedings of the 6th International Conference on Enterprise Information Systems, Porto, Portugal, vol. 2, pp. 469–475. INSTICC Press (2004)
2. Dunteman, G.H.: Principal Components Analysis. Sage Publications Inc., USA (1989)
3. Cormack, R.M.: A Review of Classifications. JRSS, A 134, 321–367 (1971)
4. Moreira, T.B.S.: Financial and exchange crises in Asia in 1997-1998. Unb, Brasília, Brazil (2001) (in Portuguese)
5. Echeverría, R.: Ontologia del Lenguaje, 4th edn. Dolmen, Santiago, Chile (1997)
6. Flores, F.: Management and communication in the office of the future. PhD. Thesis, University of California at Berkeley (1981)
7. Flores, F.: Creando organizaciones para el futuro. Dólmen, Santiago, Chile (1996)
8. Kofman, F.: Metamanagement – The New Conscious Business. Antakarana Cultura Arte Ciência, São Paulo, Brazil (2002) (in Portuguese)

# An Image-Based Integration System for Real-Time Dispatching of Multi-robots

Li-Che Chen[1], Tien-Ruey Hsiang[2], Yu Fu[1], and Sheng-Luen Chung[1]

[1] Department of Electrical Engineering
[2] Department of Computer Science,
National Taiwan University of Science and Technology
43 Keelung Road, Section 4, Taipei, Taiwan 10607
{M9407202, trhsiang, D9407201, slchung}@mail.ntust.edu.tw

**Abstract.** The purpose of this paper is to design an image-based platform for real-time dispatching of multi-robots, in the context of, for example, supporting a courier transportation system. With this integration platform, we are able to compare efficiencies with different dispatching policies. With the global information regarding the robots and pick-up points available from the image process, this real-time dispatching platform is capable of detecting current positions of robots and their relative positions of destination of transportation, of planning online motion trajectories for each robot to reach its destination smoothly, and of maintaining collision avoidance while robots are moving. Different dispatching strategies, as a combination of pre-positioning of robots and swapping strategies of responsible zones for individual robots in transportation operation, are conducted on a miniature implementation platform to simulate corresponding performance of response time.

**Keywords:** Multi-robot, dispatching system.

## 1 Introduction

In order to study the efficiency of the multi-robot system with different dispatching policies [1], this paper presents the design and implementation of an image-based integration platform for a multi-robot system; robots are essentially semi-autonomous and a central computer distributes assignments to each robot. To dispatch multi-robots in real-time for transportation assignment, the proposed integration platform utilizes image process technique to detect the current positions of robots relative to transportation destination; to design a motion trajectory for each robot in carrying transportation assignment to its destination smoothly. In addition, collision avoidance strategy, while robots are moving, is designed based on global observation of the whole transportation system.

The rest of this paper is organized as follows. Main steps and requirement of real-time dispatching are introduced in Section II. Section III presents the integration platform, which includes software architecture with a miniature working space for transportation simulation. Simulation results with different dispatching policies,

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 491–498, 2008.

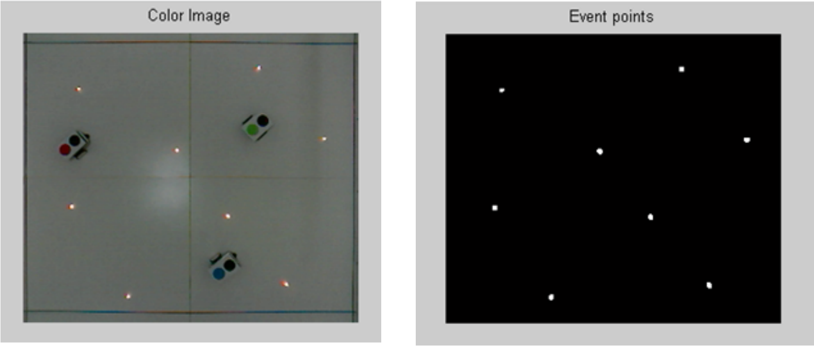
conducted on the aforementioned platform, are provided in Section IV before a brief discussion and future work given in Section V.

## 2 Real-Time Dispatching

Three processes are employed in this image-based dispatching system. The image process is to obtain global information [2] regarding the current positions of robots and requesting points. Then, a motion trajectory planning process [3][4][5] is to guide a robot to reach its assigned destination, and while doing so, with the explicit consideration given to collision avoidance [6].

### 2.1 Image Processing

The image process is to obtain global information regarding the robots and requesting points. This information is critical for later motion trajectory planning and collision avoidance. There are two procedures involved in the image process: the initial detection procedure and the subsequent robot tracking procedure. The initial detection is to obtain the initial configuration information, as shown in Fig. 1, including number, positions and identifications of the robots, as well as of requesting points in the system.



(a) (b)

**Fig. 1.** The regions of interest obtained from initial detection procedure

After the initial detection procedure, the dispatching system needs to keep track of the most update locations of robots once the dispatching operation starts. In doing so, it would be redundant to scan the whole image to update the most recent position of each robot. Instead, given the maximum speed of a robot, its next position is bound to be within a certain area. Accordingly, we develop a tracking procedure to reduce the search area of the robot, thus increasing the efficiency of image processing. Fig. 2 shows the region of interest for the detection of a known robot based on its previous detected location.

In summary, our image process works like the following: Starting with the robot's initial position obtained in the initial detection, we can predict its next position to be

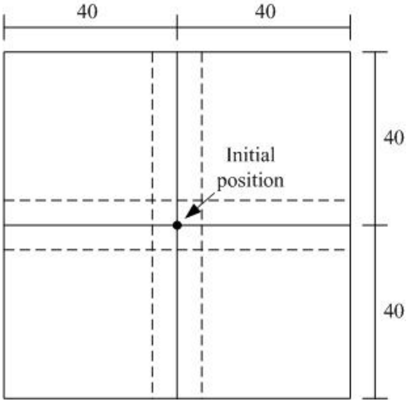


**Fig. 2.** The prediction array for robot's next position

within the 80×80 pixel array rather than the whole global image array. Then, the robot tracking procedure to obtain exact position of the target robot is the same as initial detection, with the region of interested limited to this smaller the 80×80 pixel array.

## 2.2 Motion Trajectory Planning

Dispatching decision determines which robot is to be sent to a particular requesting point (source), and following that, to another (destination) point. The process of motion trajectory planning has to do with how a robot traverses to reach a particular point in the transportation space. After the relative positions of a robot and its destination are obtained from the aforementioned image process step, motion trajectory for each dispatched robot is to be developed.

As shown in Fig. 3, let $\vec{a}$ be the vector of the front direction of the robot; $\vec{b}$, the vector of the direction from the robot to its destination. The angle between the direction of the robot and its direction of the destination can be calculated by Eq.(1).

$$\theta = \cos^{-1} \frac{\vec{a} \cdot \vec{b}}{|\vec{a}||\vec{b}|} \tag{1}$$

Sign of the result shows the direction to rotate, as indicated in equation (2),

$$\vec{a} \times \vec{b} \rightarrow \begin{cases} >0 & \quad \textit{turnleft} \\ \ 0 \ \Rightarrow & \quad \textit{hold} \\ <0 & \quad \textit{turnright} \end{cases} \tag{2}$$

With the additional relative positions of the dispatched robot and its destination being known from the image process, a motion trajectory path can be constructed, as shown in Fig. 4, where the dash line denotes the motion trajectory that the host computer will use to direct the robot to move smoothly to its destination.

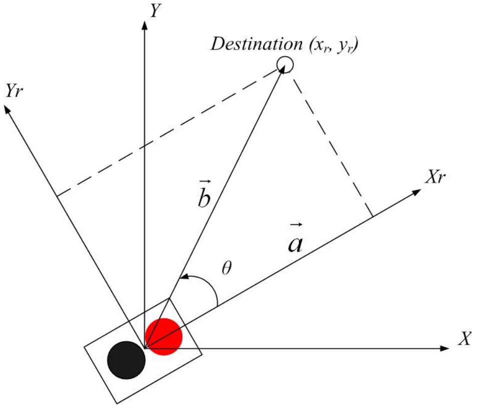


**Fig. 3.** The relative position of a robot to its destination

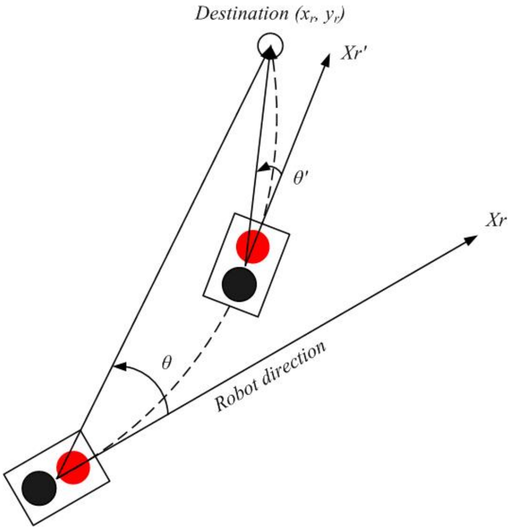


**Fig. 4.** The motion trajectory planning

Because the angle between the robot direction and the destination changes as the robot moves, modification of the motion trajectory plan is performed real-time until the robot reaches its assigned destination. In practice, motion trajectory plan is translated into "move" command, with two parameters of *orientation* and *distance*, issued at sample time interval from the host computer to the designated robot.

### 2.3 Collision Avoidance

In commanding robots move in the transportation space, special attention needs be given to collision avoidance: collision into an obstacle, or into another robot. In particular, for

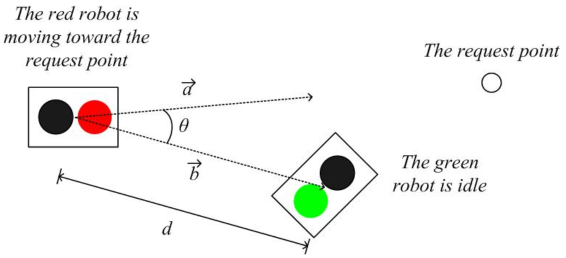


**Fig. 5.** Collision avoidance with a idle robot

the latter case of collisions into another robot, we develop two modes of collision avoidance: one is to avoid collision with an idle robot, and the other, a moving robot.

Shown in Fig. 5, suppose that a robot (red) on the left is moving toward its destination, while another robot (green) nearby stands idle. Let $\vec{a}$ denote the vector between the moving robot to its destination; $\vec{b}$, the vector from the moving robot to the idle robot. In addition, $d$ denotes the distance between these two robots, and $\theta$, the angle between the direction of the moving robot and direction of the vector from this moving robot to the idle robot, where $\theta$ is calculated by Eq.(1). If the distance $d$ is less than a certain value, indicating a potential risk of collision, then the moving robot will be directed to “detour” from the straight line that would otherwise lead into collision: move command issued to the moving robots is based on the current combination of distance $d$ and angle $\theta$, which in effect resulting in different motion trajectory designed by experience rules.

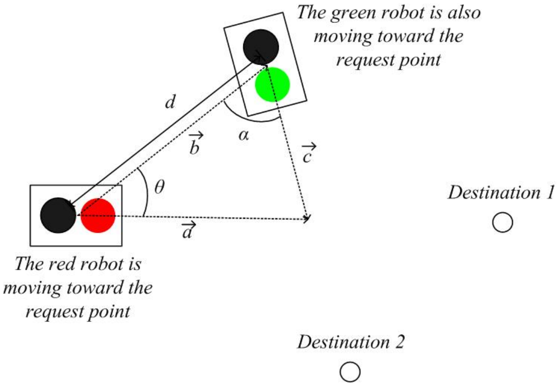

The green robot is also moving toward the request point
d
$\vec{b}$
$\alpha$
$\vec{c}$
$\theta$
Destination 1
$\vec{a}$
The red robot is moving toward the request point
Destination 2

**Fig. 6.** Collision avoidance with another moving robot

On the other hand, if the robot is moving toward another moving robot, as shown in Fig. 6, a different rule applies. Let $\vec{a}$ denote the vector of the first robot to its destination; $\vec{b}$, the vector from the first robot to the second one; $\vec{c}$, the vector of the second robot to its destination. In addition, let $d$ be the distance between these two robots; $\theta$ the angle between the direction of the first robot to its destination and the direction from the first robot to the second robot; and *α* the angle between the vector of the second robot to its destination to the direction from the second robot to the first robot. If the distance *d* is less than a certain value, then the dispatching host will initiate mode two of collision avoidance: one robot has to stop while letting the other keeps moving. The resolution rule is defined as followed: if the angle *θ* is less than *α*, meaning that the other robot is closer to the intersection of the moving trajectories of the two than it; subsequently, the robot associated with smaller angle has to stop until the other robot has moved out the potential collision region. On the contrary, if the angle *θ* is larger than *α*, then, again, the robot associated with the smaller anger has to stops till the other one clear of the potential collision region.

## 3 Integrated Platform

In order to implement the proposed method, we construct a miniature, experimental platform for multi-robot system. As shown in Fig. 7, the miniature working space is constructed in a 180×180 $cm^2$ board, with eight embedded fixed requesting points represented by LED. Each requesting point has two different color LEDs, green and red: the green one to denote a pick-up point, while the red one to denote a delivery destination point.

The image-based dispatching system comprises of several components: a central control system, Bluetooth modules, robots, a system platform and a camera. The Boe-Bot are Bluetooth-accessible, waiting to serve in the system by receiving move commands from the host computer. These robots have no sensor to locate themselves. Instead, an overhead camera, Logitech QuickCam Notebooks Pro., is installed to collect global information regarding the relative positions of the robots and request points.

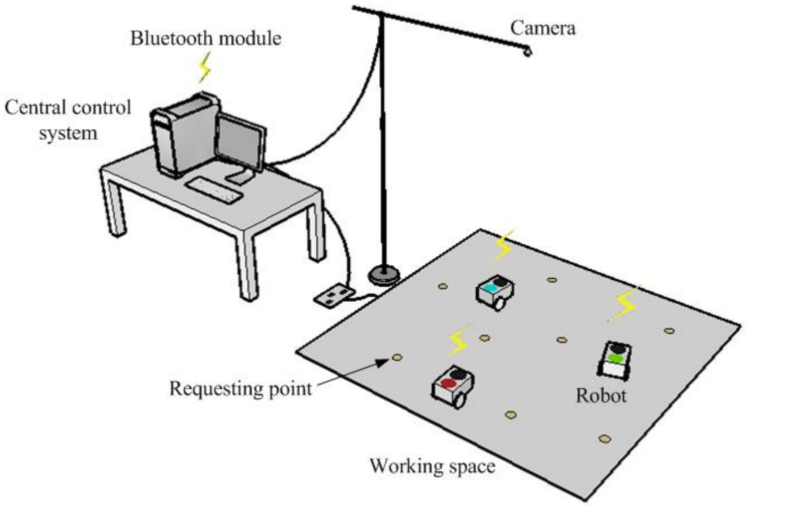


**Fig. 7.** A miniature, experimental platform

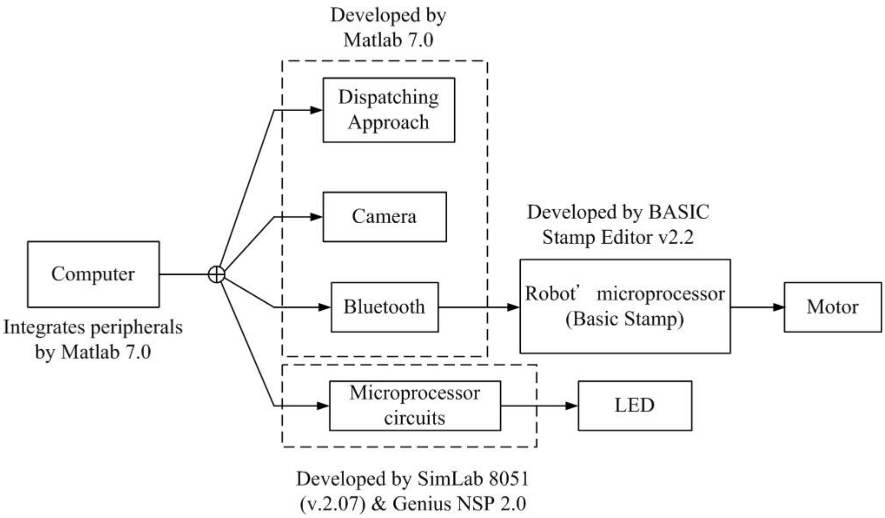

Developed by
Matlab 7.0
Dispatching Approach
Camera
Bluetooth
Computer
Integrates peripherals by Matlab 7.0
Developed by BASIC Stamp Editor v2.2
Robot' microprocessor (Basic Stamp)
Motor
Microprocessor circuits
LED
Developed by SimLab 8051 (v.2.07) & Genius NSP 2.0

**Fig. 8.** Software architecture of the image-based dispatching system

The central control system, whose software architecture is illustrated in Fig. 8, is developed by Matlab 7.0 and implemented on a Windows XP.

## 4 Simulation with Different Dispatching Strategies

The image-based integration platform for multi-robot system is designed and implemented to compare performance of different dispatching policies. Functional blocks in Fig. 9 illustrate the interactions involved when executing a dispatching policy.

Three different dispatching policies have been simulated on the miniature work space with the image-based dispatching platform: First, dispatching without pre-position and without switching; Second, dispatching with pre-position but without

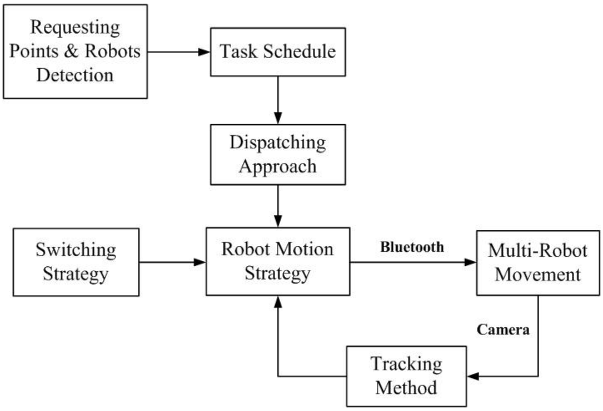


**Fig. 9.** Functional block diagram in executing a dispatching policy

switching; and the third, dispatching with both pre-position and switching. It is shown in [1] that, in terms of the response time of robots to service requests, the second and the third policies improves upon the first one by 37.78% and 54.05% respectively, thus confirming the better performance of the third dispatching policy.

## 5 Discussion

Based on the framework presented in this work, we are implementing a real-time dispatching system in a building on campus of NTUST, where several cameras, instead of one, are installed. Our final goal is to guide a platoon of robots in performing transportation requests occurred at different places in the building.

**Acknowledgement.** This work was supported in part by National Science Council, Taiwan under Grant NSC 96-2221-E-011-125 and in part by the Ministry of Education, Taiwan under Grant MOE 96-E-01-297.

## References

1. Chen, L.C.: Multi-Robot Dispatching in a Geographically Constrained Environment. M.S. thesis, Dept. Electrical Eng., National Taiwan University of Science and Technology, Taipei, Taiwan (2007)
2. Crowly, J.L., Pourraz, F.: Continuity properties of the appearance manifold for mobile robot position estimation. Image and Vision Computing 19, 741–752 (2001)
3. Defoort, M., Palos, J., Kokosy, A., Floquet, T., Perruquetti, W., Boulinguez, D.: Experimental motion planning and control for an autonomous nonholonomic mobile robot. In: IEEE International Conference on Robotics and Automation, pp. 2221–2226 (2007)
4. Latombe, J.C.: Motion planning: a journey of robots, molecules, digital actors, and other artifacts. The International Journal of Robotics Research 18(11), 1119–1128 (1999)
5. Zhang, H., Ostrowski, J.P.: Visual motion planning for mobile robots. IEEE Transactions on Robotics and Automation 18(2), 199–208 (2002)
6. Borenstein, J., Koren, Y.: Real-time obstacle avoidance for fast mobile robots. IEEE Transactions on System, Man, and Cybernetics 19(5), 1179–1187 (1989)

# A Fast Duplication Checking Algorithm for Forward Reasoning Engines

Takahiro Koh[1], Yuichi Goto[1], and Jingde Cheng[1]

Department of Information and Computer Sciences, Saitama University,
338-8570 Saitama, Japan
{t_koh,gotoh,cheng}@aise.ics.saitama-u.ac.jp

**Abstract.** A forward reasoning engine capable of relevant reasoning is an indispensable component in many advanced knowledge-based systems with purposes of discovery or prediction. Any forward reasoning engine has to deal with duplication checking of intermediate results that is the most time-consuming process in forward reasoning. Therefore, to decrease execution time of forward reasoning engines is a crucial issue for their successful applications. This paper presents a fast algorithm for duplication checking process in forward reasoning engines, analyzes its time and space complexities, and shows its effectiveness by some experimental results.

## 1 Introduction

Forward reasoning engine is a computer program to automatically draw new conclusions by repeatedly applying inference rules to given premises and obtained conclusions until some previously specified conditions are satisfied [1]. A forward reasoning engine capable of relevant reasoning is an indispensable component in many advanced knowledge-based systems with purposes of discovery or prediction [1]. Discovery is the process to find out or bring to light of that which was previously unknown. Prediction is the action to make some future event known in advance, especially on the basis of special knowledge, and therefore, it is a notion must relate to a point of time to be considered as the reference time. For any discovery/prediction, both the discovered/predicted thing and its truth must be unknown before the completion of discovery/prediction process. There is no discovery/prediction process that does not perform reasoning because reasoning is the only way to draw previously unknown new conclusions from given premises [1,2]. Relevant reasoning requires that for any argument to be valid there must be some connection of meaning, i.e., some relevance, between its premises and its conclusion, among other things [3]. Relevant reasoning must play the key role in discovery/prediction because any discovery or prediction has not an explicitly given target as its goal.

To decrease execution time of forward reasoning engines is a crucial issue for their successful applications because forward reasoning engines should get enough conclusions in acceptable time [4]. For example, some applications for

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 499–507, 2008.

the purpose of prediction need to get conclusions about some future events before the events occur [4].

FreeEnCal [1], a forward reasoning engine with general purpose, was developed in order to interpret and perform inference rules defined and given by its users, draw fragments of various classical and non-classical logic systems formalized as Hilbert style axiomatic systems, Gentzen natural deduction systems, or Gentzen sequent calculus systems, draw empirical theorems of various formal theories constructed based on various logic systems, and perform deductive, inductive, and abductive reasoning automatically. However, the current FreeEnCal is too slow because of duplication checking process, which is the most time-consuming process in forward reasoning.

In order to improve efficiency of FreeEnCal, this paper presents a fast algorithm for duplication checking process in forward reasoning engines, and shows its effectiveness by some experimental results. This paper is organized as follows. In section 2, we explain forward reasoning engine, and precisely define the duplication checking. In section 3, we present a fast duplication checking algorithm and analyzes its time and space complexities. In section 4, we show some experimental results. In section 5, we discuss about our experimental results. Concluding remarks are given in section 6.

## 2 FreeEnCal: A Forward Reasoning Engine with General Purpose

### 2.1 Overview of FreeEnCal

FreeEnCal has been developed as a forward reasoning engine with general purpose. FreeEnCal has facilities where users can configure premises and inference rules, and deduce conclusions by applying the inference rules to the premises. FreeEnCal performs the following four processes repeatedly.

**Inference rule selection process:** it selects an inference rule from inference rules previously specified by user.

**Reasoning process:** it selects required number of premises by the inference rule which is selected in inference rule selection process. The permutation of the premises and the inference rule must have never chosen. If it is possible to apply the inference rule to the premises, then deduce conclusions.

**Duplication checking process:** it finds all of the conclusions duplicated against a premise or a previously deduced conclusion.

**Adding process:** it adds all conclusions which are judged as new conclusions in duplication checking process.

An issue on the current FreeEnCal is that it takes too long execution time. One of the causes of this problem is the duplication checking algorithm because the duplication checking process is one of the most time-consuming process. However, the duplication checking process is indispensable from the viewpoint of prediction or discovery because some conclusion may be the same as a premises or other conclusion and such conclusions are not only unnecessary from the viewpoint of prediction or discovery, but also will cause performance decreasing.

### 2.2 The Naive Duplication Checking Algorithm

The duplication checking process finds all conclusions that match to at least one premise or one previously deduced conclusion. In this paper, a subject and a pattern is a logical formula. A subject is checked whether or not it matches to a pattern. A subject matches to a pattern if a subject is unifiable to the pattern by substituting to variables in the pattern.

In order to implement the duplication checking process, we should solve the following problem:

**The Duplication Checking Problem.** For a given finite set of patterns $\{p_1$ , ... , $p_n\}$ and a given finite set of subjects $\{s_1$ , ... , $s_m\}$, find all the subjects that match to at least one of the patterns.

A naive duplication checking algorithm is as follows.

**Algorithm 1.** *Naive Algorithm*

```
Procedure DuplicationCheck(ListOfPatterns,ListOfSubjects)
For each formula $s$ in ListOfSubjects begin
   For each formula $p$ in ListOfPatterns begin
      If Matching(p,s) returns true then
         Output $s$;
      end (if)
   end (for)
end (for)
```

This naive algorithm merely compares every pattern to every subject. This algorithm calls a procedure *Matching* shown in algorithm 2, which checks whether or not specified subject matches to pattern. It returns true if subject matches to pattern. Otherwise, it returns false.

**Algorithm 2.** *Matching Algorithm*

```
Procedure Matching(pattern, subject)
If pattern is sentential variable then
   If pattrn is substituted then
      If substitution of pattrn is the same logical formula as subject then
         Return true;
      end (if)
   end (if)
   substitute subject to pattern;
   Return true;
Else if main connector in pattrn and subject is the same then
   For each pair of formulas $< p_i, s_i >$ such that $p_i$ is the $i$-th airty of
pattern and $s_i$ is the $i$-th arity of subject then
      If Matching($p_i$, $s_i$) returns false then
```

```
            **Return** false;
        **End (if)**
    **End (for)**
    **Return** true;
**End (if)**
**Return** false;
```

Algorithm 2 is one to deal with process propositional formulas, but it is very easy to extend to process first order predicate formulas. The naive algorithm takes $O(N \times M \times L)$ time where $N$ is the number of patterns and $M$ is the number of subjects and $L$ is the number of characters in a pattern. The naive algorithm takes $O(L \times N + L \times M)$ space because this space is needed to hold input.

## 3 A Fast Duplication Checking Algorithm

The naive algorithm works very slow because it performs a lot of unnecessary comparisons, that is, comparing the same prefixes of logical formulas in polish notation. However, it is not necessary to compare the same prefixes more than one time. For example, two subjects "$\rightarrow \vee ABB$" and "$\rightarrow \vee AB \wedge AB$," and two patterns "$\rightarrow \wedge CDC$" and "$\rightarrow \wedge CD \vee CD$," where $\rightarrow$, $\vee$, and $\wedge$ are binary connectives and $A$, $B$, $C$ and $D$ are sentential variables. We should compare the first subject and the first pattern, but we don't need to compare anymore because the prefix "$\rightarrow \vee$" does not match to "$\rightarrow \wedge$." However, the naive algorithm compares same prefixes. The key idea of our new duplication checking algorithm is to find duplicated subjects without such unnecessary comparison.

In order to never compare same prefixes again, we adopt trie [5] to hold patterns and subjects. Trie is an ordered tree data structure, and all the descendants of any one node have a common prefix of the string associated with that node. A path from root to leaf expresses a value stored into a tree. Our algorithm takes two trees as an input, one is a list of patterns and the other is a list of subjects. Our algorithm is based on the depth first search algorithm with branch and bound, and it starts searching from each root node simultaneously. Let us consider a node $n_p$ in the tree of patterns and a node $n_s$ in the tree of subjects, and $Prefix(n)$ is the prefix associated with node $n$. It is possible apply the branch and bound method if $Prefix(n_s)$ does not match to $Prefix(n_p)$ because all the other subjects having $Prefix(n_s)$ does not match to any patterns having $Prefix(n_p)$. If it reaches to a leaf node in the tree of subjects, the subject matches to a pattern. Therefore, it should output the subject. We define our algorithm as algorithm 3. This algorithm consists from a function "DuplicationCheck". It is a recursive function and requires two arguments, *PatternNode* and *SubjectNode*. The *PatternNode* is a node in the tree of patterns, and the *SubjectNode* is a node in the tree of subjects. When this function is called, $Prefix(SubjectNode)$ should be matched to $Prefix(PatternNode)$. This function finds all the pairs of $Prefix(descendantofSubjectNode)$ and $Prefix(childofPatternNode)$ that

$Prefix(descendantofSubjectNode)$ matches to $Prefix(childofSubjectNode)$. When the pair of such a node is found, this function calls recursively itself by using these nodes as arguments. Algorithm 3 is one to deal with propositional formulas, but it is very easy to extend to process first order predicate formulas.

**Algorithm 3.** *A Fast Duplication Checking Algorithm*

```
Procedure DuplicationCheck(PatternNode,SubjectNode)
If PatternNode is a leaf node then
    Output the prefix on PatternNode;
End (if)
For each edges $p$ in PatternNode begin
    $c$ := the label of $p$;
    If $c$ is variable then
        If $c$ is substituted then
            If there is a logical formula substituted to $c$ and the same sub-formula in SubjectNode then
                DuplicationCheck(Next node of $p$,Next node of the path of the same sub-formula);
            end (if)
        else
            For each formula $f$ starting from SubjectNode begin
                Substitute $f$ to $c$;
                DuplicationCheck(Next node of $p$, Next node of $f$);
                Unbind form $c$;
            end (for)
        end (if)
    else
        If SubjectNode has edge labeled $c$ then
            $s$ := edge of SubjectNode labeled $c$;
            DuplicationCheck(Next node of $p$,Next node of $s$)
        end (if)
    end (if)
end (for)
```

Let $N$ be the number of patterns, and $M$ be the number of subject, and $L$ be the number of characters in a pattern. In this paper, we consider all the patterns consists from the same number of characters to simplify the situation. In order to analyze the time complexity, we point out to the number of comparison between characters because it is the most frequent operation in these two algorithms. In this case, the time complexity is proportional to the number of edges in trees. And space complexity is also proportional to the number of edges in trees. In the worst case for time complexity, the structure of tree is as shown in figure 1.

In this case, the number of edges in pattern's tree is $L \times N$ and the number of edges in subject's tree is $L \times M$, and our algorithm performs comparison at most $L \times M \times N$ times. Therefore, our algorithm works $O(L \times M \times N)$ time and takes $O(L \times M + L \times N)$ space. This is same as the naive algorithm. However, it is not

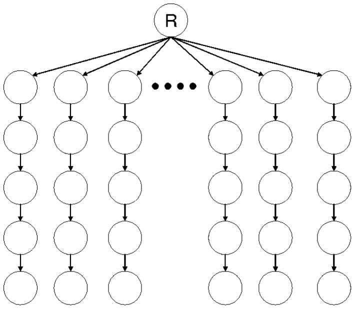


**Fig. 1.** Structure of tree in worst case

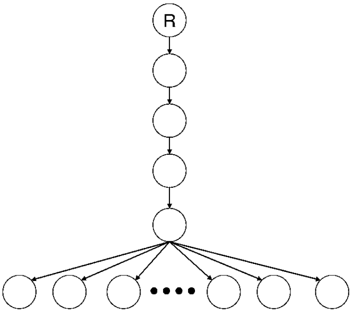


**Fig. 2.** Structure of tree in best case

serious because the kinds of characters is enough less than $N$ or $M$. Therefore, it is impossible to make such worst case input when a forward reasoning engine treats a large number of premises and conclusions In the best case for time complexity, the structure of tree is as shown in figure 2. In this case, the number of edges in pattern's tree is $L - 1 + N$ and the number of edges in subject's tree is $L - 1 + M$, and our algorithm compares at most $L - 1 + N \times M$ times. Therefore, our algorithm works $O(L + N \times M)$ time and takes $O(M + N)$ space. It is also unreal because of the same reason as the worst case. Let us consider a case that tree is perfect $k$-ary tree as an average case. In the perfect $k$-ary tree, there are

$$k \times \frac{k^L - 1}{k - 1}$$

edges. The number of leaf nodes is $k^L$. Therefore,

our algorithm compares at most

$$\frac{(M \times N - 1)}{(M^{\frac{1}{L}} \times N^{\frac{1}{L}} - 1)} \times M^{\frac{1}{L}} \times N^{\frac{1}{L}}$$

times. Our algorithm works $O(N \times M)$ time and takes $O(M + N)$ space. Therefore, our algorithm works at most $L$ times faster than the naive algorithm and there is no significant difference between our algorithm and the naive algorithm from the viewpoint of space complexity.

## 4 Experimental Results

To evaluate the performance of our algorithm in FreeEnCal, we have implemented our algorithm in FreeEnCal and have measured its execution time of the duplication checking process and the number of comparison between a character in a pattern and a sub-formula in a subject when deducing conclusions from axioms of the following some logic systems: Classical Mathmatical Logic with entailment and negation (CMLen for short), System K with enatilment, negation

**Table 1.** Limitations of Degree

| Premises | → | ∨ | ∧ | ¬ | □ |
|---|---|---|---|---|---|
| CMLen | 3 | - | - | ∞ | - |
| K | 3 | - | - | 1 | 1 |
| Re | 4 | - | - | - | - |
| Tc | 3 | 1 | 1 | ∞ | - |

**Table 2.** The number of conclusions

| Premises | Total | Non-duplicated |
|---|---|---|
| CMLen | 998,873 | 11,063 |
| K | 187,565 | 20,243 |
| Re | 454,699 | 39,258 |
| Tc | 2165,795 | 556,549 |

**Table 3.** Execution time(sec)

| Premises | Naive | Fast | ratio |
|---|---|---|---|
| CMLen | 301.29 | 0.56 | 538 |
| K | 846.24 | 1.48 | 572 |
| Re | 9,150.11 | 7.86 | 1164 |
| Tc | 739,685.52 | 18.86 | 39220 |

**Table 4.** The number of comparison

| Premises | Naive | Fast | ratio |
|---|---|---|---|
| CMLen | $1.22 \times 10^9$ | $1.73 \times 10^6$ | 705 |
| K | $2.86 \times 10^9$ | $3.90 \times 10^6$ | 733 |
| Ee | $4.92 \times 10^9$ | $8.97 \times 10^6$ | 549 |
| Tc | $1.77 \times 10^{12}$ | $4.79 \times 10^7$ | 36990 |

and box (K for short), System R [6,7] of relevant implicatoin with entailment (Re for short), and Tc [2,3] in strong relevance logics (Tc for short). We select the modus ponens as an inference rule for all premises and the necessitation for an additional inference rule for K. We select the law of double negation as an elimination rule for K and Tc. We also select the law of double modal operator as an additional elimination rule for K and the law of conjunction/disjuncion as additional elimination rules for Tc. In addition, we should limit the number of deduced conclusions to be finite. Therefore, we limit by degree of nest of a logical connector [8,9]. The limitations of degree are shown in table 1. We have experimented on a computer exclusively. Table 2 shows the number of deduced conclusions and the number of conclusions that is not duplicated.

Table 3 shows the execution time and its speed up ratio. The execution time of our algorithm is about 300 to 39,000 times shorter than the naive algorithm. Table 4 shows the number of comparison and its decreasing ratio. The number of comparison in our algorithm is about 360 to 37,000 times smaller than the naive one. From these tables, our algorithm is effective when the number of formulas in the input increases.

Table 5 shows the maximum size of process' data segement of FreeEnCal. From this table, FreeEnCal with our algorithm consumes more amount of memory than FreeEnCal with naive algorithm only in a case of using axioms in Tc as premises.

**Table 5.** Used memory

| Premises | Naive | Fast | ratio |
|---|---|---|---|
| CMLen | 155.4 | 104.5 | 0.672 |
| K | 130.3 | 111.6 | 0.865 |
| Ee | 519.6 | 317.7 | 0.611 |
| Tc | 1,090.4 | 1,715.6 | 1.573 |

## 5 Discussion

As we analyze in section 3, the time complexity of our algorithm is at most $L$ times faster than the naive algorithm from the viewpoint of time complexity. Our algorithm works fast if pruning occurs frequently. Therefoere, our algorithm works fast if the number of matched subjects is less because the pruning occurs frequently. A performance of our algorithm depends on data. From our experimental results, execution time is shortened when the number of comparison decreases. Therefore, it is possible to consider that there is a correlation between them. Hence, our algorithm is faster than the naive algorithm because the number of comparison in our algorithm is at most the same as the naive one, and works clearly faster than the naive algorithm in our experiment.

Our algorithm may work efficiently in a forward reasoning engine based on strong relevant logics because a tree may become a good shape, that is, few branches occur near root node in the tree. In strong relevant logics, some logical formulas in classical mathematical logics is not allowed because the kinds of variables on logical formula in strong relevant logics is restricted by strong relevance [3,7], the kinds of variables is at most the half of classical mathematical logics. Therefore, branches near its root node may not occur in tree of strong relevant logics.

Our algorithm does not depend on FreeEnCal because our algorithm is possible to process any logical formulas in polish notation. Therefore, our algorithm is able to be applied to any forward reasoning engines to decrease its execution time.

From the viewpoint of space complexity, there is no significant difference between our algorithm and the naive algorithm becuase both algorithm takes $O(L \times N + L \times M)$ space in the worst case. Moreover, there is no significant difference on the maximum size of process' data segment in our experiment. On the other hands, FreeEnCal with our algorithm consumes more amount of memory than FreeEnCal with naive algorithm only in a case of using axioms in Tc as premises. It is caused by a bad shape of tree. In this case, there are 4 logical connectives. Therefore, branches near its root node occur more frequently than other cases that there are less number of logical connectives.

## 6 Concluding Remarks

We have presented a fast duplication checking algorithm for forward reasoning engines, analyzed its time and space complexities, and shown its effectiveness by some experimental results.

As a related work, the Rete algorithm [10] is an efficient pattern matching algorithm for implementing production rule systems. The Rete algorithm stores partial matches as cache data and avoids complete re-evaluation of all facts each time changes are made to the production system's working memory. Our algorithm does not store cache data. Moreover, the Rete algorithm cannot be used to solve the duplication checking problem.

We are developing a fast algorithm for the reasoning process, another time consuming process in forward reasoning engine, to decrease its execution time.

## References

1. Cheng, J., Nara, S., Goto, Y.: FreeEnCal: A Forward Reasoning Engine with General-Purpose. In: Apolloni, B., Howlett, R.J., Jain, L. (eds.) KES 2007, Part II. LNCS (LNAI), vol. 4693, pp. 444–452. Springer, Heidelberg (2007)
2. Cheng, J.: A strong relevant logic model of epistemic processes in scientific discovery. In: Kawaguchi, E., Kangassalo, H., Jaakkola, H., Hamid, I.A. (eds.) Information Modelling and Knowledge Bases XI. Frontiers in Artificial Intelligence and Applications, vol. 61, pp. 136–159. IOS Press, Amsterdam (2000)
3. Cheng, J.: Strong Relevant Logic as the Universal Basis of Various Applied Logics for Knowledge Representation and Reasoning. Information Modelling and Knowledge Bases XVII, 310–320 (2006)
4. Goto, Y., Nara, S., Cheng, J.: Efficient anticipatory reasoning for anticipatory systems with requirements of high reliability and high security. International Journal of Computing Anticipatory Systems 14, 156–171 (2004)
5. Fredkin, E.: Trie memory. Commun. ACM 3(9), 490–499 (1960)
6. Anderson, A.R., Belnap, N.D.: Entailment: The Logic of Relevance and Necessity, vol. 1. Princeton University Press, Princeton (1975)
7. Anderson, A.R., Belnap, N.D., Dunn, J.: Entailment: The Logic of Relevance and Necessity, vol. 2. Princeton University Press, Princeton (1992)
8. Cheng, J.: EnCal: An Automated Forward Deduction System for General-Purpose Entailment Calculus. In: Proceedings of IFIP World Conference on Advanced, pp. 507–517 (September 1996)
9. Nara, S., Omi, T., Goto, Y., Cheng, J.: A general-purpose forward deduction engine for modal logics. In: Knowledge-Based Intelligent Information and Engineering Systems, 9th International Conference, pp. 739–745 (September 2005)
10. Forgy, C.L.: Rete: a fast algorithm for the many pattern/many object pattern match problem. Expert systems: a software methodology for modern applications, 324–341 (1990)

# Formal Definition of Relevant Logical Symbol Occurrence

Noriaki Yoshiura

Department of Information and Computer Science, Saitama University
255, Shimo-ookubo, Sakura-ku, Saitama City, Saitama Prefecture, Japan
Tel.: +81-48-858-3498; Fax: +81-48-858-3498
yoshiura@fmx.ics.saitama-u.ac.jp

**Abstract.** The formalization of human reasoning is a main issue in artificial intelligence. Although relevant logic is one of most useful ways of such formalization, inference of logic does not correspond to that of human reasoning because the aim of logical inference is to deal with the truth values of propositions while that of human reasoning is to discover useful information. We introduce the concept of logical symbol occurrence relevance into relevant logic for formalization of human reasoning. This paper discusses relevance of logical symbol occurrence and gives formal definition of relevance from the discussion. Moreover, we propose new logic where all logical symbol occurrences in theorems are relevant with respect to the relevance definition of this paper.

## 1 Introduction

Relevant logic has been studied from the viewpoint of philosophy[1]. The aim of relevant logic is to remove the fallacies of implication from classical logic. The formalization of human reasoning is a main issue in artificial intelligence. As a method of formalization of knowledge reasoning, relevant logic is more suitable than classical logic in the sense that the fallacies of implication are removed[2]. Although relevant logic is one of most useful ways of such formalization, inference of logic does not correspond to that of human reasoning because the aim of logical inference is to deal with the truth values of propositions while that of human reasoning is to discover useful information. In this paper, we introduce the concept of logical symbol occurrence relevance into relevant logic for formalization of human reasoning. We use relevant logic $ER$, which is free from fallacies of implication and has more provability than relevant logic $R$ [4,5]. In $ER$, when we introduce logical symbol such as conjunction, disjunction, negation or atomic proposition into proof, we do not consider relevance of logical symbol occurrence. On the other hand, in knowledge base reasoning, when we introduce logical symbol into proof, we consider such relevance. For example, in $ER$ or relevant logic $R$, $A \wedge C \rightarrow B \wedge D$ can be inferred from $A \rightarrow B$ and $C \rightarrow D$. However, there is no guarantee that the conjunctive connection between $A$ and $C$ is relevant. We consider more concrete example; from two propositions ”It rains” → ”Picnic is canceled” and ”Ken is human” → ”Ken will die at some future”, we can infer the following proposition.

(”It rains” ∧ ”Ken is human”) → (”Picnic is canceled” ∧ ”Ken will die at some future”)

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 508–516, 2008.

However, from the viewpoint of human reasoning, there is no relation between ”It rains” and ”Ken is human”. This inference does not seem to obtain useful information. On the other hand, $A \rightarrow B \wedge C$ can be inferred from $A \rightarrow B$ and $A \rightarrow C$. In this case, we think that relevance of conjunctive connection of $B$ and $C$ is guaranteed by the same proposition $A$, which is a premise of each implication formula.

From this discussion, it is necessary to define relevance of logical symbol occurrence for inference of useful information from knowledge base. In [4], relevance of logical connective introduced by inference rules is discussed, however, relevance of all logical symbol occurrences is not dealt with. In this paper, we define relevance of logical symbol occurrence from the viewpoint of proof structure. We use proof structure of relevant logic *ER*. We think that relevance of logical symbol occurrence depends on proof structure, that is to say, how to introduce logical symbol and relationship between formulas. This paper discusses relevance of logical symbol occurrence by using proof structure.

## 2 Relevant Logic *ER*

Relevant logic *ER* is defined as a sequent style natural deduction system as showed in Fig.1[4]. The different point from usual natural deduction is that *ER* is a kind of labeled deduction system[3]. A formula has an attribute value, which shows how the formula is inferred and which kind of rule can be applied to the formula. By using attribute values, we restrict the applicability of inference rules in order to remove the fallacies of implication from *ER*.

**Definition 1.** *Atomic propositions are formulas of ER. If A and B are formulas of ER, then* $\neg A$, $A \wedge B$, $A \wedge B$, $A \vee B$ *and* $A \rightarrow B$ *are formulas of ER.*

*”e”, ”i” and ”r” are defined to be attribute values of formula. ”e” indicates that a formula with such an attribute value can be major premise of elimination rules.*

$$A : e \vdash A : e \qquad \neg A : r \vdash \neg A : r \quad \text{Axiom} \quad \frac{\Gamma, \neg A : r \vdash}{\Gamma \vdash A : e}\ RAA \quad \frac{\Gamma \vdash A : \varphi_1 \quad \Delta \vdash B : \varphi_2}{\Gamma, \Delta \vdash A \wedge B : i}\ \wedge I$$

$$\frac{\Gamma \vdash A \wedge B : e}{\Gamma \vdash A : e}\ \wedge E1 \quad \frac{\Gamma \vdash A \wedge B : e}{\Gamma \vdash B : e}\ \wedge E2 \quad \frac{\Gamma \vdash A : \varphi}{\Gamma \vdash A \vee B : i}\ \vee I1 \quad \frac{\Gamma \vdash B : \varphi}{\Gamma \vdash A \vee B : i}\ \vee I2$$

$$\frac{\Gamma \vdash A \vee B : e \quad \Delta_1, A : e \vdash C : e \quad \Delta_2, B : e \vdash C : e}{\Gamma, \Delta_1, \Delta_2 \vdash C : e}\ \vee E1 \quad \frac{\Gamma \vdash A \vee B : e \quad \Delta_1, A : e \vdash C : i \quad \Delta_2, B : e \vdash C : \varphi}{\Gamma, \Delta_1, \Delta_2 \vdash C : i}\ \vee E2$$

$$\frac{\Gamma \vdash A \vee B : e \quad \Delta_1, A : e \vdash C : \varphi \quad \Delta_2, B : e \vdash C : i}{\Gamma, \Delta_1, \Delta_2 \vdash C : i}\ \vee E3 \quad \frac{\Gamma \vdash A \vee B : e \quad \Delta_1, A : e \vdash \quad \Delta_2, B : e \vdash}{\Gamma, \Delta_1, \Delta_2 \vdash}\ \vee E4$$

$$\frac{\Gamma, A : e \vdash B : \varphi}{\Gamma \vdash A \rightarrow B : i} \rightarrow I \quad \frac{\Gamma \vdash A \rightarrow B : e \quad \Delta \vdash A : \varphi}{\Gamma, \Delta \vdash B : e} \rightarrow E \quad \frac{\Gamma \vdash A : \varphi \quad \Delta \vdash \neg A : e}{\Gamma, \Delta \vdash}\ \neg E1 \quad \frac{\Gamma \vdash A : e \quad \neg A : r \vdash \neg A : r}{\Gamma, \neg A : r \vdash}\ \neg E2$$

$$\frac{\Gamma, A : e \vdash}{\Gamma \vdash \neg A : i}\ \neg I \quad \frac{\Gamma, A : \varphi, A : \varphi \vdash}{\Gamma, A : \varphi \vdash}\ C1 \quad \frac{\Gamma, A : \varphi, A : \varphi \vdash B : \phi}{\Gamma, A : \varphi \vdash B : \phi}\ C2 \quad \frac{\Gamma, \neg A : e, \neg A : r \vdash}{\Gamma, \neg A : r \vdash}\ C3$$

$$\frac{\Gamma, \neg A : e \vdash B : \varphi}{\Gamma \vdash A \vee B : i}\ EM1 \quad \frac{\Gamma, \neg B : e \vdash A : \varphi}{\Gamma \vdash A \vee B : i}\ EM2 \quad \frac{\Gamma \vdash A \vee B : e \quad \Delta, A : e \vdash}{\Gamma, \Delta \vdash B : e}\ DS1 \quad \frac{\Gamma \vdash A \vee B : e \quad \Delta, B : e \vdash}{\Gamma, \Delta \vdash A : e}\ DS2$$

**Fig. 1.** Inference rules of *ER*

*”i” indicates that a formula with such an attribute value can not be major premise of elimination rules. ”r” indicates that a formula with such an attribute value can be discharged by the rule RAA.*

In the following, $\varphi$, $\varphi_1$, $\cdots$, $\phi$,$\phi_1$, $\cdots$ are used as meta variables of attribute value. $A : \varphi$ is called *attribute formula* where $A$ is a formula and $\varphi$ is an attribute value.

**Definition 2.** *$\Gamma \vdash \mathscr{A}$ is called sequent where $\Gamma$ is a multiset of attribute formulas and $\mathscr{A}$ is an attribute formula or an empty formula.*

*$A$ is a theorem of ER if and only if there exists a proof $F_1, \cdots, F_n$ of ER and $F_n$ is of the form $\vdash A : \varphi$. We say that $A$ is inferred from a sets $\Gamma$ of formulas if and only if $\Gamma \vdash A : \varphi$ is proved.*

*ER* has the properties such as removal of implication relevance and so on [4].

## 3 Relevance of Logical Symbol Occurrence

This section discusses relevance of logical symbol occurrence in knowledge base reasoning in *ER*. We suppose that knowledge base is a set $\mathscr{K}$ of logical formulas and introduce the following inference rule $K$ for inference on knowledge base $\mathscr{K}$ in *ER*.

$$\frac{A \in \mathscr{K}}{\vdash A : e}\ K$$

**Definition 3.** *$ER^K$ is defined to be ER with inference rule K. A proof of $ER^K$ inference rules and knowledge base $\mathscr{K}$ is defined to be a proof of $ER^K$ based on knowledge base $\mathscr{K}$. A is a theorem of $ER^K$ based on knowledge base $\mathscr{K}$ if and only if $\vdash A : \varphi$ is proved in $ER^K$ based on knowledge base $\mathscr{K}$.*

By using a proof of $ER^K$, we discuss relevance of logical symbol occurrence such as atomic proposition and logical connectives. In the following, we use $p,q,r,s$ as atomic propositions and omit attribute value if it is clear.

### 3.1 Relevance of Atomic Proposition Occurrence

We discuss relevance of atomic proposition occurrence. We think that all logical symbol occurrences are relevant in all formulas of knowledge base $\mathscr{K}$. Therefore, in this paper, atomic proposition occurrence by inference rule $K$ is relevant. We discuss atomic proposition occurrence by several proofs.

We suppose that knowledge base $\mathscr{K} = \{p, q \rightarrow s, s \rightarrow r\}$. In the left proof of Fig.2, $(p \rightarrow q \wedge r) \rightarrow s$ is inferred. $p \rightarrow q \wedge r \vdash p \rightarrow q \wedge r$ is introduced by Axiom in *ER* and $q$ is inferred by $\wedge E1$ rule. Thus, even if $r$ can be another atomic proposition, $s$ can be inferred by $\rightarrow E$. It follows that occurrence of atomic proposition $r$ is not relevant. We discuss other occurrences of atomic proposition; occurrence of $p$ in $p \rightarrow q \wedge r \vdash p \rightarrow q \wedge r$ is relevant because in $\rightarrow E$, this occurrence of $p$ is used with the occurrence of

$$\dfrac{\dfrac{q \rightarrow s \in \mathscr{K}}{\vdash q \rightarrow s}\,K \qquad \dfrac{\dfrac{p \rightarrow q \wedge r \vdash p \rightarrow q \wedge r \qquad \dfrac{p \in \mathscr{K}}{\vdash p}\,K}{p \rightarrow q \wedge r \vdash q \wedge r}\rightarrow E}{p \rightarrow q \wedge r \vdash q}\wedge E1}{\dfrac{p \rightarrow q \wedge r \vdash s}{\vdash (p \rightarrow q \wedge r) \rightarrow s}\rightarrow I}\rightarrow E \qquad \dfrac{q \vdash q}{\vdash q \rightarrow q}\rightarrow I \qquad \dfrac{\dfrac{\dfrac{s \rightarrow r \in \mathscr{K}}{\vdash s \rightarrow r}\,K \qquad \dfrac{\dfrac{q \rightarrow s \in \mathscr{K}}{\vdash q \rightarrow s}\,K \qquad q \vdash q}{q \vdash s}\rightarrow E}{q \vdash r}\rightarrow E}{\vdash q \rightarrow r}\rightarrow I$$

**Fig. 2.** Proof example 1

$p$ introduced by the rule $K$. Occurrence of $s$ in the conclusion of this proof is relevant because it is based on $q \rightarrow s$ in knowledge base $\mathscr{K}$.

We discuss the middle and right proofs of Fig.2. In the middle proof, occurrence of $q$ is not based on knowledge base or necessary. This occurrence is not relevant and any occurrence in the conclusion $q \rightarrow q$ is not relevant. In the right proof, occurrence of $q$ in $q \vdash q$ introduced by Axiom is relevant because it is used as premise of $\rightarrow E$ with $q \rightarrow s$ in $\mathscr{K}$. Occurrence of $r$ in the conclusion of proof is relevant because it is inferred from $s \rightarrow r$ in $\mathscr{K}$ by $\rightarrow E$.

### 3.2 Relevance of Logical Connective Occurrence

We discuss relevance of occurrence of conjunction, disjunction and negation from the viewpoint of connected formulas. We consider two formulas $A \wedge A$ and $A \wedge (A \vee B)$. In the first formula, the same formulas are connected by conjunction. In the second formula, $A$ and $A \vee B$ are connected by conjunction, however, $A$ can deduces $A \vee B$ trivially and $A \wedge (A \vee B)$ is the same as $A$. For example, since in daily speech, "It rains" and "It rains" is the same as "It rains", such an occurrence of conjunction is meaningless and not relevant.

These examples show that $A \wedge B$ is redundant if $A(B)$ can infer $B(A)$. It follows that such an occurrence of conjunction is not relevant. We can propose a condition for connecting two formulas by conjunction; if two formulas are connected by conjunction, then one formula can not infer the other formula. This condition is also necessary for disjunction.

With respect to implication such as $A \rightarrow (A \vee B)$ or $(A \wedge B) \rightarrow A$, if $C$ can include or deduce $D$, then $C \rightarrow D$ is trivial and the occurrence of implication is meaningless and not relevant because we can not obtain any useful information. For example, in daily speech, that "It rains" implies "It rains" is not used. From above discussion, we have a condition that if two formulas are connected by conjunction, disjunction or implication, then one formula can not infer the other formula. We define this condition formally.

**Definition 4.** *$A \asymp B$ holds if and only if $A : e \vdash B : \varphi$ or $B : e \vdash A : \varphi$ can not be inferred in ER. For a set of formula $\Gamma$ and a formula $A$, $\Gamma \asymp A$ holds if and only if $C : e \vdash A : \varphi$ can not be inferred where C is the conjunction formula of all elements in $\Gamma$ and there exists no formula B in $\Gamma$ such that $A : e \vdash B : \varphi$. We call $A \asymp B$ "Derivation relation formula".*

For any formulas $A$ and $B$, we can decide whether $A \asymp B$ holds because we have a decision procedure of $ER$[4].

$$\dfrac{\dfrac{\dfrac{\dfrac{p\vdash p \quad \dfrac{p\to q\in\mathcal{K}}{\vdash p\to q}\,K}{p\vdash q}\,\to E \qquad \dfrac{p\vdash p \quad \dfrac{p\to r\in\mathcal{K}}{\vdash p\to r}\,K}{p\vdash r}\,\to E}{p,p\vdash q\wedge r}\,\wedge I}{p\vdash q\wedge r}\,C2}{\vdash p\to q\wedge r}\,\to I \qquad \dfrac{\dfrac{\dfrac{\dfrac{p\vdash p \quad \dfrac{p\to q\in\mathcal{K}}{\vdash p\to q}\,K}{p\vdash q}\,\to E \qquad \dfrac{s\vdash s \quad \dfrac{s\to r\in\mathcal{K}}{\vdash s\to r}\,K}{s\vdash r}\,\to E}{p,s\vdash q\wedge r}\,\wedge I}{s\vdash p\to q\wedge r}\,\to I}{\vdash s\to(p\to q\wedge r)}\,\to I$$

**Fig. 3.** Proof example 2

We discuss relevant occurrence of conjunction and disjunction from the viewpoint of proof structure. We suppose that knowledge base $\mathcal{K} = \{p \to q, p \to r, s \to r\}$ in the proofs in Fig.3. In the left proof, $\wedge I$ connects $q$ and $r$ by using $p \vdash q$ and $p \vdash r$ as premises. In this case, $p$ is in left side of two sequents and this occurrence seems to guarantee the relationship between $q$ and $r$. On the other hand, in the right proof of Fig.3, the premises of $\wedge I$ are $p \vdash q$ and $s \vdash r$ and we can not find the same proposition in the left side of sequents. Thus, we can not find the relationship between $q$ and $r$ and the occurrence of conjunction in $q \wedge r$ is not relevant. This discussion concludes that two formulas should have some relationship if they are connected by conjunction. This conclusion also holds in the case of disjunction.

In $ER$, $EM1$ and $EM2$ introduce disjunction. These rules work for "exclusive middle" and, for example, $EM1$ infers $A \vee B$ from the inference of $B$ from $\neg A$. Therefore, we consider that disjunction introduced by $EM1$ or $EM2$ is relevant.

$$\dfrac{\dfrac{\begin{matrix}\vdots\\ p,p\vdash q\end{matrix}}{p\vdash q}\,C2}{\vdash p\to q}\,\to I \qquad\qquad \dfrac{\dfrac{\begin{matrix}\vdots\\ p,p\vdash q\end{matrix}}{p\vdash p\to q}\,\to I}{\vdash p\to(p\to q)}\,\to I$$

**Fig. 4.** Proof example 3

In the left proof in Fig.4, $p \to q$ is inferred, and in the right proof in Fig.4, $p \to (p \to q)$ is inferred. The above part of these proofs is the same, however, usage of $C2$ makes a difference between two proofs. For example, in daily speech, "It rains" → ("It rains" → "Athletic meeting defers") is not relevant. Cause of irrelevance seems repetition of "It rains". Therefore, such repetition of premise of implication formula does damage to relevance of implication occurrence introduced by $\to I$. Relevance of negation occurrence is discussed like implication.

### 3.3 Formal Definition of Relevance of Logical Symbol Occurrence

From the previous discussion, this subsection gives formal definition of relevance of logical symbol occurrence. We describe length of formula $A$ by $l(A)$ and define that $\langle A, n\rangle$ represents n-th occurrence logical symbol. We define relevance of logical symbol occurrence by using proofs of $ER^K$. For any inference rule in $ER$, a logical symbol occurrence $q$ in the premise of the inference rule is source of a logical symbol occurrence $p$ in the conclusion of the inference rule if and only if $p$ is based on $q$.

**Definition 5.** *We recursively define relevance of logical symbol occurrence which appears in a proof of $ER^K$ based on knowledge base $\mathscr{K}$.*

1. *Logical symbol occurrence in formula A which is introduced by inference rule K is relevant.*
2. *Logical symbol occurrence p is relevant if and only if source of p is relevant.*
3. *In the sequent $A : \varphi \vdash A : \varphi$ which is introduced by Axiom, n-th logical symbol occurrence in A in the left side is relevant if and only if n-th logical symbol occurrence in A in the right side is relevant.*
4. *Occurrence of conjunction introduced by $\wedge I$ is relevant if $A \asymp B$ holds and $\Gamma$ and $\Delta$ are the same multiset.*
5. *In $\vee E1$, $\vee E2$, $\vee E3$ or $\vee E4$, $\langle A \vee B, n \rangle$ is relevant if and only if $\langle A, n \rangle$ is relevant where $1 \leq n \leq l(A)$. $\langle A \vee B, n + l(A) + 1 \rangle$ is relevant if and only if $\langle B, n \rangle$ is relevant where $1 \leq n \leq l(B)$. Occurrence of "$\vee$" in $A \vee B$ introduced by these rules is relevant if $A \asymp B$ holds and $\Delta_1$ and $\Delta_2$ are the same multiset.*
6. *Occurrence of implication introduced by $\rightarrow I$ is relevant if $\Gamma \asymp A$ and $A \asymp B$ hold.*
7. *In $\rightarrow E$, $\langle A \rightarrow B, n \rangle$ is relevant if and only if $\langle A, n \rangle$ is relevant where $1 \leq n \leq l(A)$.*
8. *Occurrence of negation introduced by $\neg I$ is relevant if $\Gamma \asymp A$ holds and the empty right side of $\Gamma, A : e \vdash$ is relevant.*
9. *In $\neg E1$ or $\neg E2$, the empty right side of the conclusion sequent is relevant if and only if negation occurrence in $\neg A$ is relevant.*
10. *Occurrence of disjunction introduced by $EM1$ or $EM2$ is relevant if and only if $A \asymp B$ holds.*
11. *In $DS1$, $\langle A \vee B, n \rangle$ is relevant if and only if $\langle A, n \rangle$ is relevant where $1 \leq n \leq l(A)$. In $DS2$, $\langle A \vee B, n + l(A) + 1 \rangle$ is relevant if and only if $\langle B, n \rangle$ is relevant where $1 \leq n \leq l(B)$.*

We explain this definition by several examples. In the left proof of Fig.2, occurrences of an atomic proposition $r$ and conjunction in the conclusion are not relevant. In the middle proof of Fig.2, two occurrences of $q$ are not relevant. In the right proof of Fig.3, occurrence of conjunction in the conclusion is not relevant. In the right proof of Fig.4, occurrence of implication in the conclusion is not relevant. Logical symbol occurrences in the conclusions of the other example proofs are relevant.

## 4 Logic *LRO*

This section defines *LRO* (Logic of Relevant Logical Symbol Occurrence) where all logical symbol occurrences in a theorem are relevant from the view of the definition in previous section. For definition of *LRO*, we give several definitions.

**Definition 6.** *A formula of LRO is defined to be a formula of ER. A mark of LRO is defined to be a variable or constant "$\circ$". We define that $p_\alpha$ is a mark logical symbol where p is a logical symbol (logical connectives or atomic propositions) and $\alpha$ is a mark. Mark formula is a formula consisting of mark logical symbols. Mark attribute formula is a mark formula with attribute value of ER.*

$$\vdash \mathscr{A} : e \;\; \mathrm{K}\,(\mathscr{A} \in \mathscr{K}^M)$$

$$p_y : e \vdash p_y : e \;\; \text{Axiom } (p \text{ is an atomic proposition}) \qquad p_y : r \vdash p_y : r \;\; \text{Axiom } (p \text{ is an atomic proposition})$$

$$\frac{\mathscr{A} : e \vdash \mathscr{A} : e \quad \mathscr{B} : e \vdash \mathscr{B} : e}{\mathscr{A} \wedge_x \mathscr{B} : e \vdash \mathscr{A} \wedge_x \mathscr{B} : e}\, A\wedge \qquad \frac{\mathscr{A} : e \vdash \mathscr{A} : e \quad \mathscr{B} : e \vdash \mathscr{B} : e}{\mathscr{A} \vee_x \mathscr{B} : e \vdash \mathscr{A} \vee_x \mathscr{B} : e}\, A\vee$$

$$\frac{\mathscr{A} : e \vdash \mathscr{A} : e \quad \mathscr{B} : e \vdash \mathscr{B} : e \quad \mathscr{A} \asymp \mathscr{B}}{\mathscr{A} \to_\circ \mathscr{B} : e \vdash \mathscr{A} \to_\circ \mathscr{B} : e}\, A \to 1 \qquad \frac{\mathscr{A} : e \vdash \mathscr{A} : e \quad \mathscr{B} : e \vdash \mathscr{B} : e}{\mathscr{A} \to_x \mathscr{B} : e \vdash \mathscr{A} \to_x \mathscr{B} : e}\, A \to 2$$

$$\frac{\mathscr{A} : e \vdash \mathscr{A} : e}{\neg_x \mathscr{A} : e \vdash \neg_x \mathscr{A} : e}\, A\neg 1 \qquad \frac{\mathscr{A} : e \vdash \mathscr{A} : e}{\neg_x \mathscr{A} : r \vdash \neg_x \mathscr{A} : r}\, A\neg 2 \qquad \frac{\Gamma, \neg_x \mathscr{A} : r \vdash_y}{\Gamma \vdash \mathscr{A} : e}\, RAA$$

$$\frac{\Gamma \vdash \mathscr{A} : \varphi_1 \quad \Delta \vdash \mathscr{B} : \varphi_2}{\Gamma, \Delta \vdash \mathscr{A} \wedge_x \mathscr{B} : i}\, \wedge I1 \qquad \frac{\Gamma_1 \vdash \mathscr{A} : \varphi_1 \quad \Gamma_2 \vdash \mathscr{B} : \varphi_2 \quad \mathscr{A} \asymp \mathscr{B} \quad \Gamma_1 \equiv \Gamma_2}{\Gamma_1, \Gamma_2 \vdash \mathscr{A} \wedge_\circ \mathscr{B} : i}\, \wedge I2$$

$$\frac{\Gamma \vdash \mathscr{A} \wedge_x \mathscr{B} : e}{\Gamma \vdash \mathscr{A} : e}\, \wedge E1 \qquad \frac{\Gamma \vdash \mathscr{A} \wedge_x \mathscr{B} : e}{\Gamma \vdash \mathscr{B} : e}\, \wedge E2 \qquad \frac{\Gamma \vdash \mathscr{A} : \varphi}{\Gamma \vdash \mathscr{A} \vee_x \mathscr{B} : i}\, \vee I1 \qquad \frac{\Gamma \vdash \mathscr{B} : \varphi}{\Gamma \vdash \mathscr{A} \vee_x \mathscr{B} : i}\, \vee I2$$

$$\frac{\Gamma \vdash \mathscr{A} \vee_x \mathscr{B} : e \quad \Delta_1, \mathscr{A}' : e \vdash \mathscr{C} : \varphi_1 \quad \Delta_2, \mathscr{B}' : e \vdash \mathscr{C}' : \varphi_2 \quad \mathscr{A} \equiv \mathscr{A}' \quad \mathscr{B} \equiv \mathscr{B}' \quad \mathscr{C} \equiv \mathscr{C}'}{(\Gamma, \Delta_1, \Delta_2 \vdash \mathscr{C} : \rho(\varphi_1, \varphi_2)) * \tau(eq(\mathscr{A}, \mathscr{A}') \cup eq(\mathscr{B}, \mathscr{B}') \cup eq(\mathscr{C}, \mathscr{C}'))}\, \vee E1$$

$$\frac{\Gamma \vdash \mathscr{A} \vee_x \mathscr{B} : e \quad \Delta_1, \mathscr{A}' : e \vdash \mathscr{C} : \varphi_1 \quad \Delta_2, \mathscr{B}' : e \vdash \mathscr{C}' : \varphi_2 \quad \Delta_1 \equiv \Delta_2 \quad \mathscr{A} \equiv \mathscr{A}' \quad \mathscr{B} \equiv \mathscr{B}' \quad \mathscr{C} \equiv \mathscr{C}' \quad \mathscr{A} \asymp \mathscr{B}}{(\Gamma, \Delta_1, \Delta_2 \vdash \mathscr{C} : \rho(\varphi_1, \varphi_2)) * \tau(eq(\mathscr{A}, \mathscr{A}') \cup eq(\mathscr{B}, \mathscr{B}') \cup eq(\mathscr{C}, \mathscr{C}') \cup \{x = \circ\})}\, \vee E2$$

$$\frac{\Gamma \vdash \mathscr{A} \vee_x \mathscr{B} : e \quad \Delta_1, \mathscr{A}' : e \vdash \quad \Delta_2, \mathscr{B}' : e \vdash \quad \mathscr{A} \equiv \mathscr{A}' \quad \mathscr{B} \equiv \mathscr{B}'}{(\Gamma, \Delta_1, \Delta_2 \vdash) * \tau(eq(\mathscr{A}, \mathscr{A}') \cup eq(\mathscr{B}, \mathscr{B}'))}\, \vee E3$$

$$\frac{\Gamma \vdash \mathscr{A} \vee_x \mathscr{B} : e \quad \Delta_1, \mathscr{A}' : e \vdash \quad \Delta_2, \mathscr{B}' : e \vdash \quad \Delta_1 \equiv \Delta_2 \quad \mathscr{A} \equiv \mathscr{A}' \quad \mathscr{B} \equiv \mathscr{B}' \quad \mathscr{A} \asymp \mathscr{B}}{(\Gamma, \Delta_1, \Delta_2 \vdash) * \tau(eq(\mathscr{A}, \mathscr{A}') \cup eq(\mathscr{B}, \mathscr{B}') \cup \{x = \circ\})}\, \vee E4$$

$$\frac{\Gamma, \mathscr{A} : e \vdash \mathscr{B} : \varphi \quad \Gamma \asymp \mathscr{A} \quad \mathscr{A} \asymp \mathscr{B}}{\Gamma \vdash \mathscr{A} \to_\circ \mathscr{B} : i} \to I1 \qquad \frac{\Gamma, \mathscr{A} : e \vdash \mathscr{B} : \varphi}{\Gamma \vdash \mathscr{A} \to_x \mathscr{B} : i} \to I2$$

$$\frac{\Gamma \vdash \mathscr{A} \to_x \mathscr{B} : e \quad \Delta \vdash \mathscr{A}' : \varphi \quad \mathscr{A} \equiv \mathscr{A}'}{(\Gamma, \Delta \vdash \mathscr{B} : e) * \tau(eq(\mathscr{A}, \mathscr{A}'))} \to E \qquad \frac{\Gamma, \mathscr{A} : e \vdash_x \quad \Gamma \asymp \mathscr{A}}{\Gamma \vdash \neg_x \mathscr{A} : i}\, \neg I1 \qquad \frac{\Gamma, \mathscr{A} : e \vdash_x}{\Gamma \vdash \neg_x \mathscr{A} : i}\, \neg I2$$

$$\frac{\Gamma \vdash \mathscr{A} : \varphi \quad \Delta \vdash \neg_x \mathscr{A}' : e \quad \mathscr{A} \equiv \mathscr{A}'}{(\Gamma, \Delta \vdash_x) * \tau(eq(\mathscr{A}, \mathscr{A}'))}\, \neg E1 \qquad \frac{\Gamma \vdash \mathscr{A} : e \quad \neg_x \mathscr{A}' : r \vdash \neg_x \mathscr{A}' : r \quad \mathscr{A} \equiv \mathscr{A}'}{(\Gamma, \neg \mathscr{A} : r \vdash_x) * \tau(eq(\mathscr{A}, \mathscr{A}'))}\, \neg E2$$

$$\frac{\Gamma, \neg_x \mathscr{A} : e \vdash \mathscr{B} : \varphi \quad \mathscr{A} \asymp \mathscr{B}}{\Gamma \vdash \mathscr{A} \vee_\circ \mathscr{B} : i}\, EM1 \qquad \frac{\Gamma, \neg_x \mathscr{B} : e \vdash \mathscr{A} : \varphi \quad \mathscr{A} \asymp \mathscr{B}}{\Gamma \vdash \mathscr{A} \vee_\circ \mathscr{B} : i}\, EM2$$

$$\frac{\Gamma \vdash \mathscr{A} \vee_x \mathscr{B} : e \quad \Delta, \mathscr{A}' : e \vdash \quad \mathscr{A} \equiv \mathscr{A}'}{(\Gamma, \Delta \vdash \mathscr{B} : e) * \tau(eq(\mathscr{A}, \mathscr{A}'))}\, DS1 \qquad \frac{\Gamma \vdash \mathscr{A} \vee_x \mathscr{B} : e \quad \Delta, \mathscr{B}' : e \vdash \quad \mathscr{B} \equiv \mathscr{B}'}{(\Gamma, \Delta \vdash \mathscr{A} : e) * \tau(eq(\mathscr{A}, \mathscr{A}'))}\, DS2$$

$$\frac{\Gamma, \mathscr{A} : \varphi, \mathscr{A}' : \varphi \vdash_x \quad \mathscr{A} \equiv \mathscr{A}'}{(\Gamma, \mathscr{A} : \varphi \vdash_x) * \tau(eq(\mathscr{A}, \mathscr{A}'))}\, C1 \qquad \frac{\Gamma, \mathscr{A} : \varphi, \mathscr{A}' : \varphi \vdash \mathscr{B} : \phi \quad \mathscr{A} \equiv \mathscr{A}'}{(\Gamma, \mathscr{A} : \varphi \vdash \mathscr{B} : \phi) * \tau(eq(\mathscr{A}, \mathscr{A}'))}\, C2 \qquad \frac{\Gamma, \mathscr{A} : e, \mathscr{A}' : r \vdash_x \quad \mathscr{A} \equiv \mathscr{A}'}{(\Gamma, \mathscr{A} : r \vdash) * \tau(eq(\mathscr{A}, \mathscr{A}'))}\, C3$$

**Fig. 5.** The inference rules of *LRO*

**Definition 7.** *Suppose that $\mathscr{A}$ and $\mathscr{A}'$ are mark formulas. $\mathscr{A} \equiv \mathscr{A}'$ holds if and only if the formulas obtained by removing all marks from $\mathscr{A}$ and $\mathscr{A}'$ are the same. In the case of multiset of mark attribute formulas $\Gamma_1$ and $\Gamma_2$, $\Gamma_1 \equiv \Gamma_2$ is defined similarly. We call $\mathscr{A} \equiv \mathscr{A}'$ "mark equivalent formula".*

**Definition 8.** *By $l(\mathscr{A})$, we describe length of mark formula $\mathscr{A}$, which is defined to be the number of mark logical symbols in $\mathscr{A}$. $\langle \mathscr{A}, n \rangle$ is defined to be n-th occurrence mark logical symbol. Suppose that A and B are formulas obtained by removing all marks from mark formulas $\mathscr{A}$ and $\mathscr{B}$. We define that $\mathscr{A} \asymp \mathscr{B}$ holds if and only if $A \asymp B$ holds. We also define $\Gamma \asymp \mathscr{A}$ similarly. we call $\mathscr{A} \asymp \mathscr{B}$ "Derivation relation formula".*

**Definition 9.** *For mark formulas $\mathscr{A}$ and $\mathscr{A}'$ such that $\mathscr{A} \equiv \mathscr{A}'$, $eq(\mathscr{A},\mathscr{A}')$ is defined to be a minimum set of mark equations satisfying the following condition: let $\alpha$ be the mark of $\langle \mathscr{A}, n\rangle$ and $\beta$ be the mark of $\langle \mathscr{A}', n\rangle$ where $1 \leq n \leq l(\mathscr{A})$. $\alpha = \beta \in eq(\mathscr{A},\mathscr{A}')$.*

*Suppose that E is a set of mark equations. $\tau(E)$ is defined to be a minimum set of substitution satisfying the following condition: if $\alpha = \beta$ can be inferred from E, then there exists a variable or constant($\circ$) such that $\gamma/\alpha$*[1] *and $\gamma/\beta$ is in $\tau(E)$.*

**Definition 10.** *Suppose that $\Gamma$ is a multiset of mark attribute formulas and that $\mathscr{A}$ is a mark attribute formula or a mark. $\Gamma \vdash \mathscr{A}$ is defined to be sequent of LRO. $S * \mathscr{E}$ represents application of a set of substitutions $\mathscr{E}$ to sequent S.*

**Definition 11.** *Let $\varphi$ and $\phi$ be attribute values. We define function $\rho(\varphi,\phi)$ as follows: if $\varphi = i$ or $\phi = i$, then $\rho(\varphi,\phi) = i$. If $\varphi = e$ and $\phi = e$, then $\rho(\varphi,\phi) = e$. Otherwise, $\rho(\varphi,\phi) = r$.*

**Definition 12.** *For knowledge base $\mathscr{K}$, $\mathscr{K}^M$ is a set of mark formulas obtained by attaching mark "$\circ$" to all logical symbol occurrences of formulas in $\mathscr{K}$.*

**Definition 13.** *In LRO, a proof based on knowledge base $\mathscr{K}$ is defined to be a finite sequence $F_1, \cdots, F_n$ of sequent, derivation relation formula, mark equivalence formula or set inclusion relation. $F_1, \cdots, F_n$ also must satisfy the following conditions.*

1. *If $F_i$ is a sequent, then $F_i$ is an axiom in Fig.5 or $F_i$ is inferred by one of the inference rules in Fig.5 by using $F_1, \cdots F_{i-1}$ as premises.*
2. *In the inference rule K in Fig.5, knowledge base $\mathscr{K}$ is used.*
3. *In the inference rules of Fig.5, if more than one sequents are used as premise, then the same mark except $\circ$ does not exist in more than one sequents.*
4. *In $A\wedge$, $A\wedge$, $A \rightarrow 1$, $A \rightarrow 2$, $A\neg 1$, $A\neg 2$, $\wedge I1$, $\vee I1$, $\vee I2$ and $\rightarrow I2$, the new mark introduced by these inference rules does not exist in the premises of them.*

*In LRO, the sequent F is provable if and only if there exists a proof $F_1, \cdots, F_n F$.*

**Definition 14.** *Suppose that in LRO, $\vdash \mathscr{A} : \varphi$ is proved in knowledge base $\mathscr{K}$ and that all marks in $\mathscr{A}$ are "$\circ$". Let A be a formula obtained by removing all marks from $\mathscr{A}$. A is defined to be a theorem based on knowledge base $\mathscr{K}$ in LRO.*

## 5 Conclusion

This paper discusses relevance of logical symbol occurrences by using relevant logic *ER*. We give formal definition of relevance of logical symbol occurrences and logic *LRO* whose inference is based on knowledge base. Knowledge base decides relevance of logical symbol occurrences and *LRO* guarantees that all logical symbol occurrences in theorems are relevant. We plan to give semantics of *LRO* as future work.

## References

1. Anderson, Belnap: Entailment. The logic of relevance and necessity. Princeton University Press, Princeton (1975)
2. Cheng, J.: The Fundamental Role of Entailment in Knowledge Representation and Reasoning. Journal of Computing and Information 2(1), 853–873 (1996)

[1] $\alpha/\beta$ stands for substitution $\alpha$ into $\beta$.

3. Gabbay, D.M.: Labelled Deductive Systems. Oxford Logic Guides 33, vol. 1. Oxford U.P., Oxford (1996)
4. Yoshiura, N., Yonezaki, N.: A Decision Procedure for the Relevant Logic ER, Tableaux 2000. Published in University of St. Andrews Research Report CS/00/01, pp.109–123 (2000)
5. Yoshiura, N., Yonezaki, N.: Provability of Relevant Logic ER. In: Information modeling and knowledge bases XIII, pp. 100–114. IOS Press, Amsterdam (2001)
6. Yoshiura, N.: Logic of Relevant Connectives for Knowledge Base Reasoning. In: Information modeling and knowledge bases XIV, pp. 66–80. IOS Press, Amsterdam (2002)

# Deontic Relevant Logic as the Logical Basis for Representing and Reasoning about Legal Knowledge in Legal Information Systems

Jingde Cheng

Department of Information and Computer Sciences, Saitama University,
338-8570 Saitama, Japan
cheng@ics.saitama-u.ac.jp

**Abstract.** To represent and reason about various laws, legal rules, and precedents in legal information systems, we need a right fundamental logic system to provide us with a logical validity criterion of legal reasoning as well as a formal representation language. This position paper discusses why classical mathematical logic, its classical conservative extensions, or its non-classical alternatives are not suitable candidates for the fundamental logic, and shows that deontic relevant logic is a more hopeful candidate for the fundamental logic we need.

**Keywords:** Legal reasoning, Knowledge representation, Knowledge management, Knowledge discovery, Deontic logic, Relevant logic.

## 1 Introduction

To represent, specify, verify, reason about, and ensure various laws, legal rules, and precedents in legal information systems, to make legal decisions based on legal information systems, and to discover new legal knowledge from legal information systems, we need a right fundamental logic system to provide us with a logical validity criterion of legal reasoning as well as a formal representation and specification language. The question, “Which is the right logic?” invites the immediate counter-question “Right for what?” Only if we certainly know what we need, we can make a good choice. It is obvious that different applications may require different characteristics of logic.

Because making legal decisions based on legal information systems and discovering legal knowledge from legal information systems are concerned incomplete or sometime even inconsistent laws, legal rules, and precedents, the fundamental logic must be able to underlie truth-preserving and relevant reasoning in the sense of conditional, ampliative reasoning, paracomplete reasoning, paraconsistent reasoning, and normative reasoning. This position paper discusses why classical mathematical logic, its classical conservatives extensions, or its non-classical alternatives are not suitable candidates for the fundamental logic, and shows that deontic relevant logic is a more hopeful candidate for the fundamental logic we need.

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 517–525, 2008.

## 2 Basic Notions

We firstly present some basic notions that are necessary to the discussion and claims of this paper. In fact, many problems in literature are caused by mis-definition and misunderstanding of basic notions about reasoning, entailment, and logic.

***Reasoning*** is the ***process*** of drawing ***new conclusions*** from given premises, which are already known facts or previously assumed hypotheses to provide some ***evidence*** for the conclusions. Therefore, reasoning is intrinsically ampliative, i.e., it has the function of enlarging or extending some things, or adding to what is already known or assumed. In general, a reasoning consists of a number of arguments ***in some order***. An ***argument*** is a set of ***statements*** (or ***declarative sentences***) of which one statement is intended as the ***conclusion***, and one or more statements, called "***premises***," are intended to provide some evidence for the conclusion. An argument is a conclusion standing in relation to its supporting evidence. In an argument, a claim is being made that there is some sort of ***evidential relation*** between its premises and its conclusion: the conclusion is supposed to ***follow from*** the premises, or equivalently, the premises are supposed to ***entail*** the conclusion. Therefore, the correctness of an argument is a matter of the ***connection*** between its premises and its conclusion, and concerns the ***strength*** of the relation between them. Thus, what is the criterion by which one can decide whether the conclusion of an argument or a reasoning really does follow from its premises or not? It is logic that deals with the validity of argument and reasoning in a general theory.

A ***logically valid reasoning*** is a reasoning such that its arguments are justified based on some ***logical validity criterion*** provided by a logic system in order to obtain correct conclusions. Today, there are so many different logic systems motivated by various philosophical considerations. As a result, a reasoning may be valid on one logical validity criterion but invalid on another. For example, the ***classical account of validity***, which is one of fundamental principles and assumptions underlying classical mathematical logic and its various conservative extensions, is defined in terms of ***truth-preservation*** (in some certain sense of truth) as: an argument is valid if and only if it is impossible for all its premises to be true while its conclusion is false. Therefore, a classically valid reasoning must be ***truth-preserving***. On the other hand, for any correct argument in scientific reasoning as well as our everyday reasoning, its premises must somehow be ***relevant*** to its conclusion, and vice versa. The ***relevant account of validity*** is defined in terms of ***relevance*** as: for an argument to be valid there must be some connection of meaning, i.e., some relevance, between its premises and its conclusion. Obviously, the relevance between the premises and conclusion of an argument is not accounted for by the classical logical validity criterion, and therefore, a classically valid reasoning is not necessarily relevant.

***Proving*** is the process of finding a justification for an explicitly ***specified statement*** from given ***premises***, which are already known facts or previously assumed hypotheses to provide some ***evidence*** for the specified statement. A ***proof*** is a description of a found justification. A ***logically valid proving*** is a proving such that it is justified based on some logical validity criterion provided by a logic system in order to obtain a correct proof. The most intrinsic difference between reasoning and proving is that the former is intrinsically prescriptive and predictive while the latter is intrinsically descriptive and non-predictive. The purpose of reasoning is to find some new conclusion previously unknown or unrecognized, while the purpose of proving is to find a justification for some specified statement previously given. Proving has an explicitly given target as its goal while reasoning does not. Unfortunately, until now, many studies in Computer Science and Artificial Intelligence disciplines still confuse proving with reasoning.

***Logic*** deals with ***what entails what*** or ***what follows from what***, and aims at determining which are the correct conclusions of a given set of premises, i.e., to determine which arguments are valid. Therefore, the most essential and central concept in logic is the ***logical consequence relation*** that relates a given set of premises to those conclusions, which validly follow from the premises. To define a logical consequence relation is nothing else but to provide a logical validity criterion by which one can decide whether the conclusion of an argument or a reasoning really does follow from its premises or not. Moreover, to answer the question what is the correct conclusion of given premises, we have to answer the question: correct for what? Based on different philosophical motivations, one can define various logical consequence relations and therefore establish various logic systems.

In logic, a sentence in the form of 'if ... then ...' is usually called a ***conditional proposition*** or simply ***conditional*** which states that there exists a relation of sufficient condition between the 'if' part and the 'then' part of the sentence. In general, a conditional must concern two parts which are connected by the connective 'if ... then ...' and called the ***antecedent*** and the ***consequent*** of that conditional, respectively. The truth of a conditional depends not only on the truth of its antecedent and consequent but also, and more essentially, on a necessarily relevant and conditional relation between them. The notion of conditional plays the most essential role in reasoning because any reasoning form must invoke it, and therefore, it is historically always the most important subject studied in logic and is regarded as the heart of logic [1].

When we study and use logic, the notion of conditional may appear in both the ***object logic*** (i.e., the logic we are studying) and the ***meta-logic*** (i.e., the logic we are using to study the object logic). In the object logic, there usually is a connective in its formal language to represent the notion of conditional, and the notion of conditional, usually represented by a meta-linguistic symbol, is also used for representing a logical consequence relation in its proof theory or model theory. On the other hand, in the meta-logic, the notion of conditional, usually in the form of natural language, is used for defining various meta-notions and describing various meta-theorems about the object logic.

From the viewpoint of object logic, there are two classes of conditionals. One class is empirical conditionals and the other class is logical conditionals. For a logic, a conditional is called an ***empirical conditional*** of the logic if its truth-value, in the sense of that logic, depends on the contents of its antecedent and consequent and therefore cannot be determined only by its abstract form (i.e., from the viewpoint of that logic, the relevant relation between the antecedent and the consequent of that conditional is regarded to be empirical); a conditional is called a ***logical conditional*** of the logic if its truth-value, in the sense of that logic, depends only on its abstract form but not on the contents of its antecedent and consequent, and therefore, it is considered to be universally true or false (i.e., from the viewpoint of that logic, the relevant relation between the antecedent and the consequent of that conditional is regarded to be logical). A logical conditional that is considered to be universally true, in the sense of that logic, is also called an ***entailment*** of that logic.

## 3 Logic Basis for Legal Knowledge Representation and Reasoning in Legal Information Systems

The present author considers that the following four requirements are essential to the fundamental logic to underlie representing and reasoning about legal knowledge in legal information systems. First, as a general logical criterion for the validity of reasoning, the logic must be able to underlie relevant reasoning as well as truth-preserving reasoning in the sense of conditional, i.e., for any reasoning based on the logic to be valid, if its premises are true in the sense of conditional, then its conclusion must be relevant (to the premises) and true in the sense of conditional. Second, the logic must be able to underlie ampliative reasoning in the sense that the truth of conclusion of the reasoning should be recognized after the completion of the reasoning process but not be invoked in deciding the truth of premises of the reasoning. From the viewpoint to regard reasoning as the process of drawing new conclusions from given premises, any meaningful reasoning must be ampliative but not circular and/or tautological. Third, the logic must be able to underlie paracomplete reasoning and paraconsistent reasoning. In particular, the so-called principle of Explosion that everything follows from a contradiction should not be accepted by the logic as a valid principle. In general, our knowledge about a domain may be incomplete and/or inconsistent in many ways, i.e., it gives us no evidence for deciding the truth of either a proposition or its negation, and/or it directly or indirectly includes some contradictions. Therefore, reasoning with incomplete and/or inconsistent knowledge is the rule rather than the exception in our everyday lives and all scientific disciplines. Finally, because the laws and legal rules often describe only those ideal situations, when they be used to in actual situations, we need to distinguish between what ought to be done and what is the case. Therefore, a formalisation of normative notions is necessary, i.e., logic must be able to underlie normative reasoning.

Almost all current approaches to legal knowledge representation and reasoning as well as legal information systems are somehow based on classical mathematical logic its various classical conservatives extensions, or its non-classical alternatives [3, 8, 9, 11, 13-17, 21].

Classical mathematical logic (**CML** for short) was established in order to provide formal languages for describing the structures with which mathematicians work, and the methods of proof available to them; its principal aim is a precise and adequate understanding of the notion of mathematical proof. **CML** was established based on a number of fundamental assumptions. Among them, the most essential one is the classical account of validity that is the logical validity criterion of **CML** by which one can decide whether the conclusion of an argument follows from its premises or not in the framework of **CML**. However, because the relevance between the premises and conclusion of an argument is not accounted for by the classical validity criterion, a reasoning based on **CML** is not necessarily relevant. On the other hand, in **CML** the notion of conditional, which is intrinsically intensional but not truth-functional, is represented by the notion of material implication, which is intrinsically an extensional truth-function. This leads to the problem of ‘implicational paradoxes’ [1, 2, 7] as well as the problem that a reasoning based on **CML** must be circular and/or tautological but not ampliative. Moreover, because **CML** accepts the principle of Explosion, reasoning under inconsistency is impossible within the framework of **CML**. Note that the above three facts are also true to those classical conservative extensions or non-classical alternatives of **CML** where the classical account of validity is adopted as the logical validity criterion and the conditional is directly or indirectly represented by the material implication. Finally, as a tool to describe ideal mathematical proofs, **CML** does not distinguish between ideal states and actual states and cannot to underlie normative reasoning. Therefore, **CML** cannot satisfy any of the four essential requirements for the fundamental logic to underlie representing and reasoning about legal knowledge in legal information systems.

As a result, for any legal information systems based on **CML**, its various classical conservatives extensions, or its non-classical alternatives, we cannot expect that any conclusion deduced from the system must be relevant to the given premises, even if all of the premises are already known laws, legal rules, and precedents provided enough legal evidence; we also cannot expect to discover new knowledge from the system by ampliative reasoning; we also cannot expect that the system can work well under presence of inconsistency because once the system includes some inconsistency directly or indirectly, anything can be deduced from the system.

Deontic logic is a branch of philosophical logic to deal with normative notions such as obligation (ought), permission (permitted), and prohibition (may not) for underlying normative reasoning [4, 10, 12, 19, 20]. Informally, it can also be considered as a logic to reason about ideal versus actual states or behaviour. It seems to be an adequate tool to represent and reason about legal knowledge. In fact, classical deontic logic has been used in representation of laws and formalisation of legal rules [11, 15]. However, because any classical deontic logic is a classical conservatives extension of **CML**, all problems in **CML** caused by the classical account of validity and the material implication also remained in the logic. Moreover, there is the problem of deontic paradoxes in classical deontic logic [4, 10, 18].

Until now, the only family of logic adopting the relevant account of validity as the logical validity criterion is the family of relevant (relevance) logic including strong relevant (relevance) logic [1, 2, 5, 7]. A major characteristic of the relevant logics is that they have a primitive intensional connective to represent the notion of (relevant) conditional and their logical theorems include no implicational paradoxes. The underlying principle of the relevant logics is the relevance principle, i.e., for any entailment provable in a relevant logic, its antecedent and consequent must share at least one propositional variable. What underlies the strong relevant logics is the strong relevance principle: for any entailment provable in a strong relevant logic, every propositional variable in the entailment occurs at least once as an antecedent part and at least once as a consequent part. In the framework of strong relevant logics, if a reasoning is valid, then both the relevance between its premises and its conclusion and the validity of its conclusion in the sense of conditional can be guaranteed in a certain sense of strong relevance. Variable-sharing is a formal notion designed to reflect the idea that there be a meaning-connection between the antecedent and consequent of an entailment. Also, since the notion of entailment is represented in all relevant logics by a primitive intensional connective but not an extensional truth-function, a reasoning based on the relevant logics is ampliative but not circular and/or tautological. Moreover, because all relevant logics reject the principle of Explosion, they can certainly underlie paraconsistent reasoning. However, the relevant logics cannot underlie normative reasoning.

Consequently, what we need is a suitable deontic extension of strong relevant logics such that it can satisfy all of the above four essential requirements. The deontic relevant logics (see appendix) are obtained by introducing deontic operators and related axiom schemata and inference rules into strong relevant logics, and they can satisfy all of the four essential requirements for the fundamental logic system to underlie representing and reasoning about legal knowledge in legal information systems.

The deontic relevant logics provide a formal language with normative notions which can be used as a formal representation and specification language for representing and specifying laws, legal rules, and precedents. The logics also provide a sound logical basis for reasoning about laws and legal rules as well as verifying them. Based on the logics, truth-preserving and relevant reasoning in the sense of conditional, ampliative reasoning, paracomplete reasoning, paraconsistent reasoning, and normative reasoning are all possible. The logics also provide a foundation for constructing more powerful logic systems to deal with other issues in legal knowledge representation and reasoning. For examples, we can add temporal operators and related axiom schemata into the logics in order to represent and reason about legal propositions and relationships among them that are time-dependent. We can also add epistemic operators and related axiom schemata into the logics in order to represent and reason about epistemic processes of lawyers and judges.

## 4 Concluding Remarks

We have shown that that deontic relevant logic is a hopeful candidate for the fundamental logic to underlie representing and reasoning about legal knowledge in legal information systems. The deontic relevant logic can satisfy all requirements from various aspects of representing, specifying, verifying, reasoning about, and discovering legal knowledge in legal information systems. To our knowledge, no other logic proposed for legal knowledge representation and reasoning has this advantage.

The propositional deontic relevant logics was first proposed by Tagawa and Cheng to solve the well-known problem of deontic paradoxes in classical deontic logic [18]. The idea to adopt deontic relevant logic as the fundamental logic to underlie representing and reasoning about legal knowledge in legal information systems was first proposed by Cheng in a short paper at ACM SAC '06 [6]. We are working on real applications of deontic relevant logics in building practical legal information systems.

## References

1. Anderson, A.R., Belnap Jr., N.D.: Entailment: The Logic of Relevance and Necessity, vol. I. Princeton University Press, Princeton (1975)
2. Anderson, A.R., Belnap Jr., N.D., Dunn, J.M.: Entailment: The Logic of Relevance and Necessity, vol. II. Princeton University Press, Princeton (1992)

3. Anderson, B.: “Discovery” in Legal Decision-Making. Kluwer Academic, Dordrecht (1996)
4. Aqvist, L.: Deontic Logic. In: Gabbay, D., Guenthner, F. (eds.) Handbook of Philosophical Logic, 2nd edn., vol. 8, pp. 147–264. Kluwer Academic, Dordrecht (2002)
5. Cheng, J.: A Strong Relevant Logic Model of Epistemic Processes in Scientific Discovery. In: Kawaguchi, E., Kangassalo, H., Jaakkola, H., Hamid, I.A. (eds.) Information Modelling and Knowledge Bases XI. Frontiers in Artificial Intelligence and Applications, vol. 61, pp. 136–159. IOS Press, Amsterdam (2000)
6. Cheng, J.: Deontic Relevant Logic as the Logical Basis for Legal Information Systems. In: Proc. 21st Annual ACM Symposium on Applied Computing, pp. 319–320. ACM Press, New York (2006)
7. Dunn, J.M., Restall, G.: Relevance Logic. In: Gabbay, D., Guenthner, F. (eds.) Handbook of Philosophical Logic, 2nd edn., vol. 6, pp. 1–128. Kluwer Academic, Dordrecht (2002)
8. Hage, J.C.: Reasoning with Rules: An Essay on Legal Reasoning and Its Underlying Logic. Kluwer Academic, Dordrecht (2002)
9. Hage, J.: Studies in Legal Logic. Springer, Heidelberg (2005)
10. Hilpinen, R.: Deontic Logic. In: Goble, L. (ed.) The Blackwell Guide to Philosophical Logic, pp. 159–182. Blackwell, Oxford (2001)
11. Jones, A.J.I., Sergot, M.: Deontic Logic in the Representation of Law: Towards a Methodology. Artificial Intelligence and Law 1(1), 45–64 (1992)
12. Nute, D. (ed.): Defeasible Deontic Logic. Kluwer Academic, Dordrecht (1997)
13. Prakken, H.: Logical Tools for Modelling Legal Argument: A Study of Defeasible Reasoning in Law. Kluwer Academic, Dordrecht (1997)
14. Prakken, H., Sartor, G.: The Role of Logic in Computational Models of Legal Argument: A Critical Survey. In: Kakas, A.C., Sadri, F. (eds.) Computational Logic: Logic Programming and Beyond. LNCS (LNAI), vol. 2408, pp. 342–381. Springer, Heidelberg (2002)
15. Royakkers, L.M.M.: Extending Deontic Logic for the Formalisation of Legal Rules. Kluwer Academic, Dordrecht (1998)
16. Stelmach, J., Brozek, B.: Methods of Legal Reasoning. Springer, Heidelberg (2006)
17. Stranieri, A., Zeleznikow, J.: Knowledge Discovery from Legal Databases. Springer, Heidelberg (2005)
18. Tagawa, T., Cheng, J.: Deontic Relevant Logic: A Strong Relevant Logic Approach to Removing Paradoxes from Deontic Logic. In: Ishizuka, M., Sattar, A. (eds.) PRICAI 2002. LNCS (LNAI), vol. 2417, pp. 39–48. Springer, Heidelberg (2002)
19. von Wright, G.H.: Deontic Logic. Mind 60, 1–15 (1951)
20. von Wright, G.H.: Norm and Action. Routledge & Kegan Paul (1963)
21. Zeleznikow, J., Hunter, D.: Building Intelligent Legal Information Systems. Kluwer Law and Taxation Publishers, Dordrecht (1994)

## Appendix. Systems of Deontic Relevant Logics

The logical connectives, deontic operators, axiom schemata, and inference rules of deontic relevant logics are as follows:

**Primitive logical connectives:**

$\Rightarrow$ (entailment), $\neg$ (negation), $\wedge$ (extensional conjunction)

**Defined logical connectives:**

$\otimes$ : intensional conjunction, $A \otimes B =_{df} \neg(A \Rightarrow \neg B)$

$\oplus$ : intensional disjunction, $A \oplus B =_{df} \neg A \Rightarrow B$

$\Leftrightarrow$: intensional equivalence, $A\Leftrightarrow B =_{df} (A\Rightarrow B)\otimes(B\Rightarrow A)$

$\vee$ : extensional disjunction, $A\vee B =_{df} \neg(\neg A\wedge\neg B)$

$\rightarrow$ : material implication, $A\rightarrow B =_{df} \neg(A\wedge\neg B)$ or $A\rightarrow B =_{df} \neg A\vee B$

$\leftrightarrow$ : extensional equivalence, $A\leftrightarrow B =_{df} (A\rightarrow B)\wedge(B\rightarrow A)$

**Deontic operators:**

$\boldsymbol{O}$ : obligation operator, $\boldsymbol{O}A$ means "It is obligatory that A"

$\boldsymbol{P}$ : permission operator, $\boldsymbol{P}A =_{df} \neg\boldsymbol{O}(\neg A)$, $\boldsymbol{P}A$ means "It is permitted that A"

**Axiom schemata:**

E1 $A\Rightarrow A$, E2 $(A\Rightarrow B)\Rightarrow((C\Rightarrow A)\Rightarrow(C\Rightarrow B))$, E2′ $(A\Rightarrow B)\Rightarrow((B\Rightarrow C)\Rightarrow(A\Rightarrow C))$

E3 $(A\Rightarrow(A\Rightarrow B))\Rightarrow(A\Rightarrow B)$, E3′ $(A\Rightarrow(B\Rightarrow C))\Rightarrow((A\Rightarrow B)\Rightarrow(A\Rightarrow C))$

E3″ $(A\Rightarrow B)\Rightarrow((A\Rightarrow(B\Rightarrow C))\Rightarrow(A\Rightarrow C))$

E4 $(A\Rightarrow((B\Rightarrow C)\Rightarrow D))\Rightarrow((B\Rightarrow C)\Rightarrow(A\Rightarrow D))$, E4′ $(A\Rightarrow B)\Rightarrow(((A\Rightarrow B)\Rightarrow C)\Rightarrow C)$

E4″ $((A\Rightarrow A)\Rightarrow B)\Rightarrow B$, E4‴ $(A\Rightarrow B)\Rightarrow((B\Rightarrow C)\Rightarrow(((A\Rightarrow C)\Rightarrow D)\Rightarrow D))$

E5 $(A\Rightarrow(B\Rightarrow C))\Rightarrow(B\Rightarrow(A\Rightarrow C))$, E5′ $A\Rightarrow((A\Rightarrow B)\Rightarrow B)$

N1 $(A\Rightarrow(\neg A))\Rightarrow(\neg A)$, N2 $(A\Rightarrow(\neg B))\Rightarrow(B\Rightarrow(\neg A))$, N3 $(\neg(\neg A))\Rightarrow A$

C1 $(A\wedge B)\Rightarrow A$, C2 $(A\wedge B)\Rightarrow B$, C3 $((A\Rightarrow B)\wedge(A\Rightarrow C))\Rightarrow(A\Rightarrow(B\wedge C))$

C4 $(\boldsymbol{L}A\wedge\boldsymbol{L}B)\Rightarrow\boldsymbol{L}(A\wedge B)$, where $\boldsymbol{L}A =_{df} (A\Rightarrow A)\Rightarrow A$

D1 $A\Rightarrow(A\vee B)$, D2 $B\Rightarrow(A\vee B)$, D3 $((A\Rightarrow C)\wedge(B\Rightarrow C))\Rightarrow((A\vee B)\Rightarrow C)$

DCD $(A\wedge(B\vee C))\Rightarrow((A\wedge B)\vee C)$, C5 $(A\wedge A)\Rightarrow A$, C6 $(A\wedge B)\Rightarrow(B\wedge A)$

C7 $((A\Rightarrow B)\wedge(B\Rightarrow C))\Rightarrow(A\Rightarrow C)$, C8 $(A\wedge(A\Rightarrow B))\Rightarrow B$, C9 $\neg(A\wedge\neg A)$, C10 $A\Rightarrow(B\Rightarrow(A\wedge B))$

DR1 $\boldsymbol{O}(A\Rightarrow B)\Rightarrow(\boldsymbol{O}A\Rightarrow\boldsymbol{O}B)$, DR2 $\boldsymbol{O}A\Rightarrow\boldsymbol{P}A$, DR3 $\neg(\boldsymbol{O}A\wedge\boldsymbol{O}\neg A)$

DR4 $\boldsymbol{O}(A\wedge B)\Rightarrow(\boldsymbol{O}A\wedge\boldsymbol{O}B)$, DR5 $\boldsymbol{P}(A\wedge B)\Rightarrow(\boldsymbol{P}A\wedge\boldsymbol{P}B)$

**Inference rules:**

$\Rightarrow$E : "from A and $A\Rightarrow B$ to infer B" (Modus Ponens)

$\wedge$I : "from A and B to infer $A\wedge B$" (Adjunction)

$\boldsymbol{O}$-necessitation : "if A is a logical theorem, then so is $\boldsymbol{O}A$" (Deontic Generalization)

Thus, various relevant logic systems may now defined as follows, where we use "A | B" to denote any choice of one from two axiom schemata A and B.

$\mathbf{T}_{\Rightarrow}$ = {E1, E2, E2′, E3 | E3″} + $\Rightarrow$E

$\mathbf{E}_{\Rightarrow}$ = {E1, E2 | E2′, E3 | E3′, E4 | E4′} + $\Rightarrow$E

$\mathbf{E}_{\Rightarrow}$ = {E2′, E3, E4″} + $\Rightarrow$E, $\mathbf{E}_{\Rightarrow}$ = {E1, E3, E4‴} + $\Rightarrow$E

$\mathbf{R}_{\Rightarrow}$ = {E1, E2 | E2′, E3 | E3′, E5 | E5′} + $\Rightarrow$E

$\mathbf{T}_{\Rightarrow,\neg} = \mathbf{T}_{\Rightarrow}$ + {N1, N2, N3}, $\mathbf{E}_{\Rightarrow,\neg} = \mathbf{E}_{\Rightarrow}$ + {N1, N2, N3}, $\mathbf{R}_{\Rightarrow,\neg} = \mathbf{R}_{\Rightarrow}$ + {N2, N3}

$\mathbf{T} = \mathbf{T}_{\Rightarrow,\neg}$ + {C1~C3, D1~D3, DCD} + $\wedge$I

$\mathbf{E} = \mathbf{E}_{\Rightarrow,\neg}$ + {C1~C4, D1~D3, DCD} + $\wedge$I

$\mathbf{R} = \mathbf{R}_{\Rightarrow,\neg}$ + {C1~C3, D1~D3, DCD} + $\wedge$I

$\mathbf{Tc} = \mathbf{T}_{\Rightarrow,\neg}$ + {C3, C5~C10}, $\mathbf{Ec} = \mathbf{E}_{\Rightarrow,\neg}$ + {C3~C10}, $\mathbf{Rc} = \mathbf{R}_{\Rightarrow,\neg}$ + {C3, C5~C10}

Here, $\mathbf{T}_{\Rightarrow}$, $\mathbf{E}_{\Rightarrow}$, and $\mathbf{R}_{\Rightarrow}$ are the purely implicational fragments of **T**, **E**, and **R**, respectively, and the relationship between $\mathbf{E}_{\Rightarrow}$ and $\mathbf{R}_{\Rightarrow}$ is known as $\mathbf{R}_{\Rightarrow} = \mathbf{E}_{\Rightarrow}+A\Rightarrow\boldsymbol{L}A$; $\mathbf{T}_{\Rightarrow,\neg}$, $\mathbf{E}_{\Rightarrow,\neg}$, and $\mathbf{R}_{\Rightarrow,\neg}$ are the implication-negation fragments of **T**, **E**, and **R**, respectively; **Tc**, **Ec**, and **Rc** are strong relevant (relevance) logics.

We can now obtain propositional deontic relevant logics as follows:

**DTc** = **Tc** + {DR1~DR5} + ***O***-necessitation

**DEc** = **Ec** + {DR1~DR5} + ***O***-necessitation

**DRc** = **Rc** + {DR1~DR5} + ***O***-necessitation

Various predicate deontic relevant logics then can be obtained by adding the following axiom schemata IQ1~IQ5 and inference rule ∀I into the propositional deontic relevant logics.

IQ1 $\forall x(A\Rightarrow B)\Rightarrow(\forall xA\Rightarrow\forall xB)$

IQ2 $(\forall xA\wedge\forall xB)\Rightarrow\forall x(A\wedge B)$

IQ3 $\forall xA\Rightarrow A[t/x]$ (if x may appear free in A and t is free for x in A, i.e., free variables of t do not occur bound in A)

IQ4 $\forall x(A\Rightarrow B)\Rightarrow(A\Rightarrow\forall xB)$ (if x does not occur free in A)

IQ5 $\forall x_1 \ldots \forall x_n (((A\Rightarrow A)\Rightarrow B)\Rightarrow B)$ $(n\geq 0)$

∀I : if A is an axiom, so is ∀xA (Generalization of axioms)

# A General Forward Reasoning Algorithm for Various Logic Systems with Different Formalizations

Yuichi Goto, Takahiro Koh, and Jingde Cheng

Department of Information and Computer Sciences
Saitama University 338-8570, Japan
{gotoh, t_koh, cheng}@aise.ics.saitama-u.ac.jp

**Abstract.** A forward reasoning engine with general-purpose should be able to deal with various logic systems with different formalizations and various formal theories based on the logic systems, and to perform deductive, inductive, and abductive reasoning based on the logic systems. This paper presents a general forward reasoning algorithm for various logic systems formalized as Hilbert style axiomatic systems, Gentzen natural deduction systems, or Gentzen sequent calculus systems, and its implementation in FreeEnCal, a forward reasoning engine with general-purpose, that we are developing.

## 1 Introduction

A forward reasoning engine is an indispensable component in many advanced knowledge-based systems with purposes of creation, discovery, or prediction. Forward reasoning engine is a computer program to automatically draw new conclusions by repeatedly applying inference rules, which are programmed in the reasoning engine or given by users to the reasoning engine as input, to given premises and obtained conclusions until some previously specified conditions are satisfied. A forward reasoning engine which can be used as a ready-made forward reasoning engine serving as a core and fundamental component in such advanced knowledge-based systems as well as an alone forward reasoning engine with general-purpose should deal with various logic systems and formal theories based on the various logic systems formalized as various formal systems and to perform deductive, inductive, and abductive reasoning.

However, there has not been such a forward reasoning engine yet. The first forward reasoning engine is "Logic Theory Machine", developed by Newell, Shaw and Simon in 1957. As a well-known fact, the Logic Theory Machine was not successful due to the problem of computational complexity [1]. This (and the resolution method discovered by Robinson) led almost all researchers to give up the approach of forward reasoning but adopt the more efficient approach of backward reasoning [2].

This paper presents an automated forward reasoning algorithm with various logic systems formalized as Hilbert style axiomatic systems, Gentzen natural

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 526–535, 2008.

deduction systems, or Gentzen sequent calculus systems, and its implementation in FreeEnCal [3], a forward reasoning engine with general-purpose, that we are developing.

The rest of this paper is organized as follows: Section 2 gives our consideration and basic idea of a general forward reasoning algorithm. Section 3 explains overview of the algorithm and presents essential parts of our algorithm. After that, Section 4 shows an implementation based the algorithm. Some concluding remarks are given in Section 5.

## 2 Basic Idea

A forward reasoning algorithm dealing with Hilbert style axiomatic systems has been already proposed and implemented in an automated forward deduction system for general-purpose, named EnCal [4,5], but the algorithm cannot deal with various logic systems because logical connectives and inference rules are preprogrammed. EnCal mainly consists of following three processes, and performs the processes repeatedly.

**Derivation process:** it checks whether an inference rule can apply to each of tuples which consist of well-formed formulas, wffs for short, given as premises and previously deduced wffs and which have not been checked yet. If the inference rule can, it deduces wffs.

**Duplication checking process:** it finds all of deduced wffs which are duplicate of given premises or previously deduced wffs.

**Adding process:** it adds all wffs which are not duplicate into a list of previously deduced wffs.

On the other hand, from the view point of syntax, the differences among a certain logic system or formal theory formalized as Hilbert style axiomatic systems, Gentzen natural deduction systems, or Gentzen sequent calculus systems, are as follows [6,7,8]; 1) Vocabulary of object logic: the main connective of sequent formula, sequent for short, is necessary for Gentzen sequent calculus systems but not for other formal systems. 2) Formulas: Hilbert style axiomatic systems and Gentzen natural deduction systems have only wffs as their formulas, but Gentzen sequent calculus systems have both wffs and sequents as its formulas. 3) Vocabulary of meta logic: variables which represent a sequence of wffs are necessary for Gentzen sequent calculus systems but not for the other formal systems. 4) Inference rules: Hilbert style axiomatic systems and Gentzen natural deduction systems have inference rules for applying a sequence of wffs, but Gentzen sequent calculus systems have inference rules for applying a sequence of sequents. Gentzen natural deduction systems have inference rules to use hypotheses, but Hilbert style axiomatic systems do not. 5) Set of initial formulas: Gentzen sequent calculus systems have a set of sequents as initial sequents, but the other two formal systems have a set of wffs as axioms. That is, the difference of Hilbert style axiom systems and Gentzen natural deduction systems is only whether it has inference rules to use hypotheses or not, and the differences of

Hilbert style axiom systems and Gentzen sequent calculus systems are whether it deals with wffs or sequents.

We can get a forward reasoning algorithm dealing with Hilbert style axiomatic systems and Gentzen natural deduction systems by introducing inference to use hypotheses into the EnCal's algorithm. A typical inference rule to use a hypothesis is like as follows.

$$\mathrm{hypo}(\psi, \varphi) \vdash \psi \rightarrow \varphi$$

$\mathrm{hypo}(\psi, \varphi)$ denotes $\psi$ is a hypothesis of $\varphi$. The inference rule means "if we can derive $\psi$ from $\varphi$, then we may conclude $\psi \rightarrow \varphi$" [7]. From the viewpoint of forward reasoning, we can consider that the inference rule means"if we can find $\psi$ on the derivation path from given premises to $\varphi$, then we may deduce $\psi \rightarrow \varphi$." In general, if a inference rule to use hypotheses is as follows,

$$\mathrm{hypo}(\psi_0, \varphi_0), \ldots, \mathrm{hypo}(\psi_n, \varphi_n) \vdash \delta_0, \ldots, \delta_m, \ (n, m \in \mathbb{N}),$$

then we can consider that the inference rule means "if we can find each of $\psi_i$ on the each derivation path from given premises to $\varphi_i$ $(0 \leq i \leq n)$ , then we may deduce $\delta_0, \ldots, \delta m$." Under such interpretation of the inference rules, it is possible to autonomously judge whether an inference rule to use hypotheses can apply to a sequence of wffs so that we can program that.

We can also get a forward reasoning algorithm dealing with Gentzen sequent calculus systems by improving the algorithm dealing with Hilbert style axiomatic systems. Most basic processes of EnCal's algorithm are unification and pattern matching between two wffs. The algorithms of unification and pattern matching are that improve algorithms proposed in [9] to deal with wffs. On the other hand, a sequent consists of equal and more than zero wffs. Unifying or matching between two sequents is unifying or matching each of wffs in a sequent and each of wffs in the other sequent. Hence, it is possible to implement unification and pattern matching processes for sequents by using unification and pattern matching processes for wffs. Moreover, it is easy to implement derivation and duplication checking process by using the unification and pattern matching process for sequents because the main work of procedures called the two processes is to control only when to call them.

## 3 A General Forward Reasoning Algorithm

### 3.1 Overview

We present a general forward reasoning algorithm dealing with the three formal systems, but there is enough space in this paper to explain the algorithm in detail. We therefore explain overview of the algorithm in this subsection. After that, we present essential parts of the algorithm following subsections, that is, derivation process for wffs, and derivation and duplication checking process for sequents, in propositional calculus.

A program based on our algorithm mainly consists of following four processes, and performs the processes repeatedly.

**Inference rule selection process:** it selects and picks an inference rule from a set of given inference rules.

**Derivation process:** it checks whether the inference rule can apply to each of tuples which consist of formulas given as premises and previously deduced formulas and which have not been checked yet. If the inference rule can, it deduces formulas.

**Duplication checking process:** it finds all of deduced formulas which are duplicate of given premises or previously deduced formulas.

**Adding process:** it adds all formulas which are not duplicate into a list of previously deduced formulas.

The program takes input data: a set of formulas as premises , a set of inference rules specified by users, and a set of natural numbers as limitations of nested logical connectives and modal operators. The formulas and inference rules are formalized as one of the three formal systems. The vocabulary of object language of this algorithm are as follows, *pattern variables* are variables which mean arbitrary logical formulas, *propositional symbols* mean propositions. *predicate symbols* mean predicates, *predicate variables* are variables which mean arbitrary predicates, *individual constants* mean names in discussion domain, *individual variables* are variables which mean arbitrary names, *sequent connective* is the main connective of sequents, *operators* means logical connectives or modal operators, *quantifiers* are universal quantifier and existential quantifier, and punctuation marks and parentheses. Definition of wffs is as follows; 1) any pattern variable is a wff, 2) any propositional symbol is a wff 3) $\rho(\tau_1, \tau_2, \cdots, \tau_i)$ is a wff where $\{\tau_i\}, (1 \leq i)$ are individual variables or constant, and $\rho$ is a predicate symbol or predicate variable, 3) $*(n, A_0, A_1, \cdots, A_n)$ is a wff where $n$ is number of wffs, $\{A_i\}, (1 \leq i)$ are wffs, and $*$ is an operator, 4) each of $\forall \xi A$ and $\exists \xi A$ is a wff where $A$ is a wff and $\xi$ is individual variable, 5) each of $\forall XA$ と $\exists XA$ is a wff where $A$ is a wff and $X$ is predicate variable, 6) nothing else is a wff. Sequents is represented as "$\Gamma \vdash \Sigma$" where $\Gamma$ is a sequence of $m$ $(\in \mathbb{N})$ wffs and $\Sigma$ is a sequence of $n$ $(\in \mathbb{N})$ wffs. Inference rules for wffs are represented as follows,

$$\mathrm{aexp}_1, \ldots, \mathrm{aexp}_m \vdash \mathrm{cexp}_1, \ldots, \mathrm{cexp}_n, \ (m, n \in \mathbb{N}).$$

$\mathrm{aexp}_i$ $(1 \leq i \leq m)$ is a wff or an expression "hypo$(A, B)$" where both $A$ and $B$ are wffs and $A$ is a hypothesis of $B$. $\mathrm{cexp}_j$ $(1 \leq j \leq n)$ is a wff. Inference rules for sequents are represented as follows,

$$\mathrm{aexp}_1, \ldots, \mathrm{aexp}_m \vdash \mathrm{cexp}, \ (m \in \mathbb{N}).$$

$\mathrm{aexp}_i$ $(1 \leq i \leq m)$ and cexp are sequents or formulas whose sequence of wffs are replaced to sequent variables. *sequent variables* are variables which mean a sequence of logical formulas, which number of wffs in the sequence is equal or more than 0. Limitations of nested logical connectives and modal operators are defined in [3].

### 3.2 Algorithm for wffs

To deal with Gentzen natural deduction systems, it is necessary to modify the derivation process of EnCal's algorithm. This subsection gives the algorithm of the modified derivation process. Derivation process can be divided two subprocesses; a process to interpret the given inference rule and a process to apply the inference rule to each sequences of wffs.

The derivation process is started by performing a procedure *Derivation*. *Derivation* takes an inference rule *IR*, an array *fList* containing all of given premises and previously deduced wffs. *Derivation* returns an array *c_pool* containing deduced new wffs.

**Algorithm 1.** *Derivation process for wffs*

```
def Derivation(IR, fList)
    initialize pre[], con[], hyp[], rec_p[], c_pool[], and bindM{}.
    InterpretIR(IR, pre, con, hyp)
    ApplyIR(fList, pre, con, hyp, rec_p, bindM, c_pool, 0, 0)
    return c_pool
end(def)
```

A procedure *InterpretIR* is called from *Derivation*. It interprets the given inference rule *IR*, and then stores wffs which occur in antecedent and consequent of the inference rule into arrays *pre*, *con*, and *hyp*.

**Algorithm 2.** *Interpretation of inference rules*

```
def InterpretIR(IR, pre, con, hyp)
    store elements of antecedent of IR into aExp
    store elements of consequent of IR into con
    pNum := number of elements in antecedent of IR
    cNum := number of elements in consequent of IR
    hNum := 0
    i := 0
    while i < pNum do
        if aExp[i] forms "hypo(A, B)" then
            hyp[i] := A
            pre[i] := B
            hNum := hNum + 1
        else
            pre[i] := aExp[i]
            hyp[i] is NULL
        end(if)
        i := i + 1
    end(while)
end(def)
```

A procedure *ApplyIR* is called from *Derivation*. It checks whether it can unify a tuple of wffs in an array *pre* and each of tuple of wffs recursively. If yes, it substitute results of the unification for each of wffs in an array *con*. *ApplyIR* takes arrays *fList*, *pre*, *con*, and *hyp*, an array *rec_p* containing the pointer to an element of *fList*, an associative array *bindM* containing results of unifications, and the natural number *pCnt* and *hCnt*.

**Algorithm 3.** *Applying a inference rule*

```
def ApplyIR(fList, pre, con, hyp, rec_p, bindM, c_pool, pCnt, hCnt)
    pNum := the number of elements in pre
    cNum := the number of elements in con
    hNum := the number of elements except empty one in hyp
    if $pCnt < pNum$ then
      index := 0
      while $index <$ the number of elements of fList do
        formula := fList[index]
        if Unify(pre[pCnt], formula, bindM) returns OK then
          $pCnt := pCnt + 1$
          if hyp[pCnt] is not NULL then
            rec_p[pCnt] := the pointer to fList[index]
          else
            rec_p[pCnt] is NULL
          end(if)
          ApplyIR(fList, pre, con, hyp, rec_p, bindM, c_pool, pCnt, hCnt)
        end(if)
        $index := index + 1$
      end(while)
    else if $hNum \neq 0$ then
      while $hCnt \leq pNum$ and hyp[hCnt] is not NULL do
        $hCnt := hCnt + 1$
      end(while)
      if $pNum < hCnt$ then
        DeduceWffs(con, bindM, c_pool)
        return END
      end(if)
      ancestors[] := GetAncestors(rec_p[hCnt])
      while (formula := pop(ancestors)) is not NULL do
          と if Unify(hyp[hCnt], formula, bindM) returns OK then
          $hCnt := hCnt + 1$
          while $hCnt \leq pNum$ and hyp[hCnt] is NULL do
            $hCnt := hCnt + 1$
          end(while)
          ApplyIR(fList, pre, con, hyp, rec_p, bindM, c_pool, pCnt, hCnt)
        end(if)
      end(while)
    else
```

```
        DeduceWffs(con, bindM, c_pool)
    end(if)
end(def)
```

A procedure *Unify*(*wff_p*, *wff_s*, *bindM*) is to check whether it can unify between *wff_p* and *wff_s* or not, according to the unification algorithm proposed in [9]. If it cannot unify then *Unify* returns 'NG', if not, then it returns 'OK' and stores results of the unification into *bindM*. A procedure *DeduceWffs*(*con*, *bindM*, *c_pool*) is to produce wffs by substituting results of unifications in *bindM* into each elements of an array *con*. After that, *DeduceWffs* adds the wffs into an array *c_pool*. A procedure *GetAncestors*(*formula*) is to return an array which contains all of wffs occurring in the derivation tree of *formula* and *formula* itself.

### 3.3 Algorithms for Sequents

To deal with Gentzen sequent calculus system, it is necessary to add derivation process and duplication process for sequents into the EnCal's algorithm.

The derivation process is started by performing a procedure *S_Derivation*. *S_Derivation* takes an inference rule *IR* and an array *fList* containing all of given sequents and previously deduced sequents. *S_Derivation* returns an array *c_pool* containing deduced new sequents.

**Algorithm 4.** *Derivation for sequents*

```
def S_Derivation(IR, fList)
    initialize pre[], con, c_pool[], and bindM{}
    store sequents in antecedent of IR into pre[]
    store a sequent in consequent of IR into con
    S_ApplyIR(fList, pre, con, bindM, c_pool, 0)
    return c_pool
end(def)
```

A procedure *S_ApplyIR* is called from *S_Derivation*. It checks whether it can unify a tuple of sequents in an array *pre* and each of tuple of sequents recursively. If yes, it substitute results of the unification for each of a sequent *con*. *S_ApplyIR* takes arrays *fList*, *pre*, and *c_pool*, a sequent *con*, an associative array *bindM* containing results of unifications, and the natural number *pCnt*.

**Algorithm 5.** *Applying a inference rule*

```
def S_ApplyIR(fList, pre, con, bindM, c_pool, pCnt)
    pNum := the number of elements in pre
    index := 0
    if $pCnt \le pNum$ then
        while $index <$ the number of elements of fList do
            formula := fList[index]
            initialize pSeqs[] and cSeqs[]
            if DelSeqVar(pre[pCnt], con, formula, pSeqs, cSeqs) returns OK then
```

```
                j := 0
                while j < the number of elements in pSeqs do
                    if S_Unify(pSeqs[j], formula, bindM) returns OK then
                        pCnt := pCnt + 1
                        S_ApplyIR(fList, pre, cSeqs[j], bindM, c_pool, pCnt)
                    end(if)
                    j := j + 1
                end(while)
            end(if)
            index := index + 1
        end(while)
    else
        DeduceSequents(c, bindM, c_pool)
    end(if)
end(def)
```

A procedure *DelSeqVars*(*premise*, *conclusion*, *formula*, *pSeqs*, *cSeqs*) is to delete all of sequent variables in *premise* and *conclusion* by substituting a sequence of wffs which occur in *formula* for, and to add the sequents gotten from *premise* and *conclusion* into an array *pSeqs* and an array *cSeqs* respectively. If *DelSeqVars* gets at least one sequent by the substitution then it returns 'OK', if not then it returns 'NG.'

A procedure *S_Unify* is to check whether it can unify *sequent_p* and *sequent_s* or not, by using the procedure *Unify* in subsection 3.2. If it cannot unify then *S_Unify* returns 'NG'. If not, then it returns 'OK' and adds results of unifications into an associative array *bindM.*

**Algorithm 6.** *Unification between two sequents*

```
def S_Unify(sequent_p, sequent_s, bindM)
    initialize pWffs[] and sWffs[]
    store all of wffs in sequent_p into pWffs by order of occurence
    store all of wffs in sequent_s into sWffs by order of occurence
    i := 0
    while i < the number of elements in pWffs do
        if Unify(pWffs[i], sWffs[i], bindM) returns NG then
            return NG
        end(if)
        i := i + 1
    end(while)
    return OK
end(def)
```

A procedure *DeduceSequents*(*conclusion*, *bindM*, *c_pool*) is to produce sequents by substituting results of unifications in an associative array *bindM* for each elements of *conclusion.* After that, it adds the sequents into *c_pool.*

The duplication checking process is started by performing a procedure *S_DuplicationChecking. S_DuplicationChecking* takes an array *fList* containing given premises and previously deduced sequents and an array *c_pool* containing deduced sequents at *S_Derivation*, and returns an array containing sequents which are not duplicate of given premises and previously deduced sequents.

**Algorithm 7.** *Duplication checking for sequents*

```
def S_DuplicationChecking(fList, c_pool)
    initialize new_pool[]
    foreach subject in c_pool
       foreach target in fList
          if S_PM(target, subject) returns NG then
             add subject into new_pool
          end(if)
       end(foreach)
    end(foreach)
    return new_pool
end(def)
```

A procedure *S_PM* is to check whether it can match *sequent_p* and *sequent_s* or not, by using the procedure *PM*. If it cannot match then *S_PM* returns 'NG'. If not, then it returns 'OK' and adds results of pattern matching into an associative array *bindM*. A procedure *PM*(*wff_p*, *wff_s*, *bindM*) is to check whether it can match *wff_p* and *wff_s* or not, according to the pattern matching algorithm proposed in [9]. If it cannot match then *PM* returns 'NG', if not, then *PM* returns 'OK' and adds results of pattern matching into an associative array *bindM*.

**Algorithm 8.** *Pattern matching between two sequents*

```
def S_PM(sequent_p, sequent_s)
    initialize pWffs[], sWffs[], bindM
    store all of wffs in sequent_p into pWffs by order of occurence
    store all of wffs in sequent_s into sWffs by order of occurence
    $i := 0$
    while $i <$ the number of elements in pWffs do
       if PM(pWffs[i], sWffs[i], bindM) returns NG then
          return NG
       end(if)
       $i := i + 1$
    end(while)
    return OK
end(def)
```

## 4 Implementation in FreeEnCal

We implemented FreeEnCal based on our algorithm and checked whether our algorithm can deal with logic systems formalized as the three formal systems.

To check our algorithm, we gives propositional classical mathematical logic system LJ formalized as Gentzen natural deduction systems and LK formalized as Gentzen sequent calculus systems [10], as input data to the FreeEnCal.

The FreeEnCal is written with C++ and complied with GCC version 4.1.3. It worked on Debian GNU/Linux Lenny in Dell PowerEdge 2850 (Main memory 4GB). FreeEnCal deduced 352 wffs when it takes a propositional symbol given as a premise, 10 inference rules of LJ, and the 3 natural numbers (3, 1, 1) as limitations of nested of material implication, conjunction, and disjunction. It deduced 10,618 sequents when it takes a sequent and 17 inference rules of LK, and the 4 natural numbers (1, 0, 0, 5) as limitations of nested of material implication, conjunction, disjunction, negation. The reason why the limitations for LK are so tight is that it is not enough memory space to deal with more loose limitations.

We therefore can consider that our algorithm can deal with logic systems formalized as the three formal systems.

## 5 Concluding Remarks

We have presented a general forward reasoning algorithm for various logic systems formalized as Hilbert style axiomatic systems, Gentzen natural deduction systems, or Gentzen sequent calculus systems, and its implementation in FreeEnCal. Some challenging issues exist: high-performance, low memory, and so on.

## References

1. Davis, M.: The Early History of Automated Deduction. In: Robinson, A., Voronkov, A. (eds.) Handbook of Automated Reasoning, vol. 1, pp. 5–15. Elsevier and MIT Press (2001)
2. Robinson, A., Voronkov, A. (eds.): Handbook of Automated Reasoning. vol. 1-2. Elsevier and MIT Press (2001)
3. Cheng, J., Nara, S., Goto, Y.: FreeEncal: A Forward Reasoning Engine with General-purpose. In: Apolloni, B., Howlett, R.J., Jain, L.C. (eds.) KES 2007, Part II. LNCS (LNAI), vol. 4693, pp. 444–452. Springer, Heidelberg (2007)
4. Cheng, J.: EnCal: An Automated Forward Deduction System for General–purpose Entailment Calculus. In: Terashima, N., Altman, E. (eds.) Advanced IT Tools, Proceedings of the 14th WCC, Canberra, pp. 507–517. Chapman & Hall, Boca Raton (1996)
5. Goto, Y., Nara, S., Cheng, J.: Efficient Anticipatory Reasoning for Anticipatory Systems with Requirements of High Reliability and High Security. International Journal of Computing Anticipatory Systems 14, 156–171 (2004)
6. Barwise, J. (ed.): Handbook of Mathematical Logic. North-Holland, Amsterdam (1977)
7. van Dalen, D.: Logic and Structure, 3rd edn. Springer, Heidelberg (1994)
8. Socher-Ambrosius, R., Johann, P.: Deduction Systems. Springer, Heidelberg (1997)
9. Winston, P.H., Horn, B.K.P.: LISP, 3rd edn. Addison-Wesley, Reading (1989)
10. Szabo, M.E. (ed.): The Collected Papers of Gerhard Gentzen. North-Holland, Amsterdam (1969)

# Map-Oriented Regional Information Management for Data Broadcasting Contents

Masahiro Ura[1], Takami Yasuda[1], Masashi Yamada[2], Mamoru Endo[2],
Shinya Miyazaki[2], and Koji Nakamura[3]

[1] Graduate School of Information Science, Nagoya University, Furo-cho, Chikusa-ku,
Nagoya-shi, Aichi, 464-8601 Japan
ura@nagoya-u.jp
[2] School of Information Science and Technology, Chukyo University
[3] Long-term Planning Division, City of Seto

**Abstract.** Although regional informatization has advanced due to the Internet, a problem remains: not all people can use such services. Data broadcasting is expected to serve as infrastructure to advance regional informatization, because it is safer and easier to use for all the people including senior citizens than internet infrastructure. We are promoting regional informatization based on data broadcasting. However, there are problems such as only information on each genre can be inspected, so the interface must be improved to present information region-wide. Moreover, published information is not managed uniformly, so maintaining the current state that recycles valuable information is difficult. In this study, we propose a map-oriented information model that focuses on maps as information interfaces.

**Keywords:** Map-oriented Information Model, Regional Informatization, Data Broadcasting.

## 1 Introduction

Although regional informatization has advanced due to the Internet, a problem remains: not all people can use such services. Data broadcasting is expected to serve as infrastructure to advance regional informatization, because it is safer and easier to use for all the people including senior citizens than internet infrastructure[1]. In the Seto City Digital Research Park Center, the delivery foothold of digital broadcasting for the Nagoya area, a structure delivers data broadcasting contents to the main public institutions in the city by regional intranet. Previously, we promoted regional informatization based on data broadcasting using this infrastructure[2]. The regional information contents produced in this project are delivered to a large-scale display in the main public facilities in the city and used as the regional population's information means. However, there are problems such as only the information on each genre can be inspected, so an interface must be improved to present information region-wide. Moreover, published information is not managed uniformly, so the current state that recycles valuable information is difficult.

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 536–542, 2008.

In this study, we propose a map-oriented information model that focuses on a map as an information interface. Regional information is expressed using this model. A map interface is used for the contents made so far, and data broadcasting contents for regional information provision are created.

## 2 Map-Oriented Information Model

Expressing information equally is difficult because it has various sides. Therefore, information management is tried from various viewpoints[3]. There is a target place in information. A map is a suitable interface from which a person may visually understand the information's position, its characteristics, and its connection. We are considering a map-oriented information model by focusing on maps as a means to express information. Information, which also changes with time, is divided into perpetual and transient elements. This model can also treat the concept of such notions of time. Information can be expressed that considers a sense of the distance of the real world by making such an information model.

### 2.1 Model Concept

In a real space, information is limited by time and location. But in virtual space, there are few restrictions, so it is more convenient, and information can be dispersed. When considering an expression method for information based on reality space, location and time are elements that can be effectively used. A map expresses the location and the characteristics of information. A correlation of the information, based on reality space, can be expressed by adding a time axis. Moreover, because information is accumulated, archives can be built based on real space.

**Location**
When information is considered based on location, it can be classified into four systems: a) specific point; b) two or more points; c) some range; d) two or more ranges (Fig. 1).

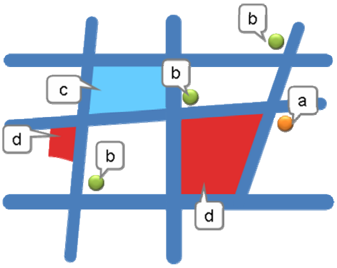


**Fig. 1.** Information on a map

**Time**
When information is considered by time, it can be classified into four systems: a) one certain point; b) some period; c) a and b joined to attain a plurality; d) something permanent without a period (Fig. 2).

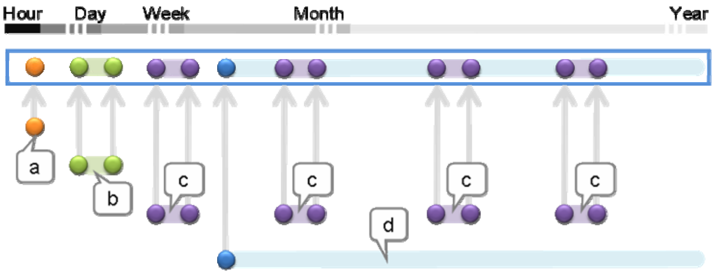


**Fig. 2.** Information on timeline

**Fusion of Location and Time**

The expression of information that considered a distance perspective in real space becomes possible by combining two concepts, "location" and "time". In other words, because the elements of time and place are added to all pieces of information, distance in a form with which people are familiar is generated between information and information. Since information is held in this form near actual space, it becomes easy for people to manipulate it.

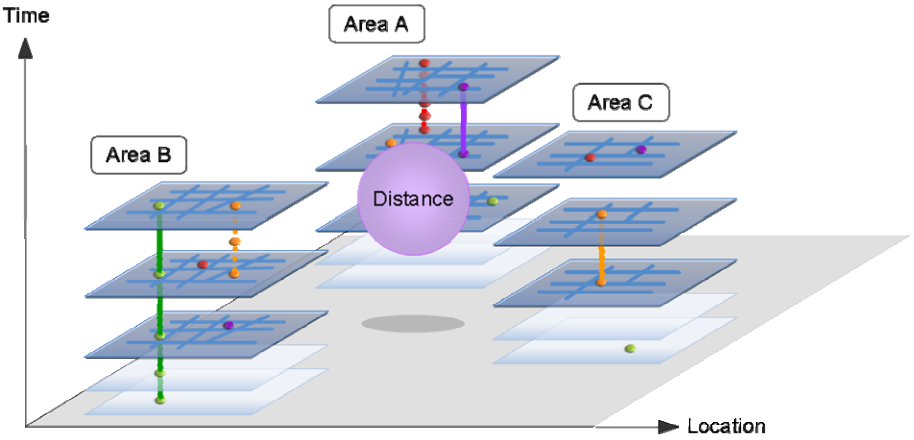


**Fig. 3.** Fusion of location and time

## 2.2 Model Structure

One example of the model based on the above concept is shown (Fig. 4). Information has the elements of position and time. These have the attributes of a point or a range, and two or more of these attributes may be added. One bit of information is composed not of the information of this model alone but the combination of such elements as heading, body text, and publisher. Map-oriented management for existing information is enabled because it adds this model. It can specify a particular point in a particular

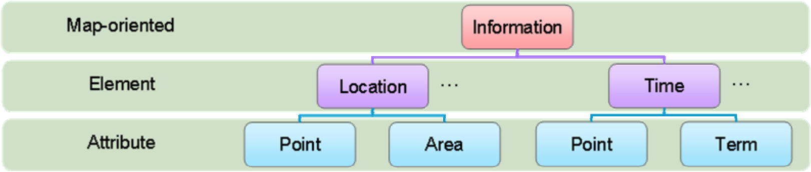


**Fig. 4.** Model node

range. Even when the time to the information and a position is ambiguous, it is related in a form with width. Thus, affinity with existing information is also high.

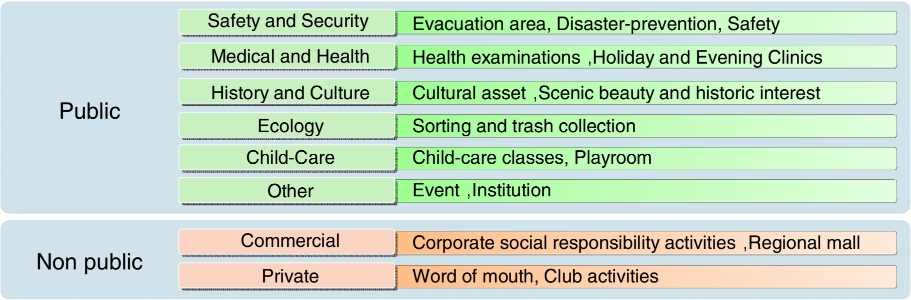
# 3 Regional Information Management

Different types of information exist in a region. The management of various regional information is enabled by applying the above map-oriented information model to such information. Moreover, because such managed information is accumulated, the database becomes the archives of the region's knowledge.

## 3.1 Regional Information

In a region, various kinds of information can be found, including public, which is possessed by municipalities and includes history and culture, and private, which is possessed by citizens (Table 1).

**Table 1.** Kinds of Regional Information



| | | |
|---|---|---|
| Public | Safety and Security | Evacuation area, Disaster-prevention, Safety |
| | Medical and Health | Health examinations ,Holiday and Evening Clinics |
| | History and Culture | Cultural asset ,Scenic beauty and historic interest |
| | Ecology | Sorting and trash collection |
| | Child-Care | Child-care classes, Playroom |
| | Other | Event ,Institution |
| Non public | Commercial | Corporate social responsibility activities ,Regional mall |
| | Private | Word of mouth, Club activities |

## 3.2 Management

As described above, a region has various kinds of information. When characteristics are divided, they become immobile things that show an institution and a place and such dynamic things as an event being held there. All of these can manage information unitarily with elements of place and time. An institution is shown as an example (Fig. 5). A museum in an area becomes an element that determines the position based on it. Moreover, the building itself is permanent. Since the museum's business hours

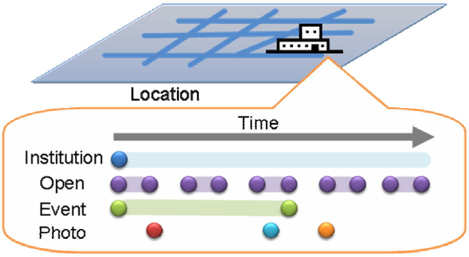


**Fig. 5.** Information of institution

are time zones that are fundamentally identical every day, they have range and continuity. Moreover, the event currently being performed has discontinuous time and a range of information. A building and the photograph of the event's scene and comments have a certain time of one point.

## 4 Data Broadcasting Contents

To confirm the effectiveness of our proposed map-oriented regional information management, we applied the data broadcasting contents produced previously. As a result, uniform management of regional information was enabled. Moreover, using the managed information, we made a page by an interface that displays a list of information on a map.

### 4.1 Data Broadcasting

In Japan, television broadcasting shifts from analog to digital in July, 2011. Data broadcasting, one of the features of digital broadcasting, can provide information as texts and images as a Web page on television. Its contents are described by Broadcast Markup Language (BML) based on XML[4].

One merit of distributing information by data broadcasting is offering information to broad layers, including such Internet non-users as the middle-aged and senior citizens. Moreover, since information can be unilaterally sent by electric waves, information can be effectively acquired during natural disasters. Thus, data broadcasting is an effective means of information distribution for public administration.

### 4.2 Contents of Regional Informatization

Previously, we tried regional informatization by producing data broadcasting contents from various viewpoints.

**Life Information**
This content is composed of the following three categories. "Disaster prevention and Safety" introduces disaster measures and shelter places. "Medical and Health" introduces holiday and evening clinics, and methods of immediate attention. "Sorting and Trash Collection" explains how to sort recyclables and provides a schedule for trash collection.

**Traditional Craft Promotion**

Seto City has been famous for ceramics for many years. Disseminating the local identity both inside and outside the region is important for planning an attractive city. This content introduces ceramic artists, knowledge concerning Seto Ceramics, the voices of regional persons engaged in Seto Ceramics, and so on.

**Tourism Promotion**

In Seto City, the citizens themselves chose 100 places to introduce the city's charm in a guide map called "One Hundred Views of Seto." This included not only places of scenic beauty and historic interest but also people's lifestyles and cultural activities. One recommended course to reflect the seasons can be experienced by Sugoroku, a game that resembles backgammon.

**Child-Care Support**

This content introduces health examinations for infants and children, health consultation counters, child-care classes, immunization programs, child-care consultation counters, playrooms, and child-care and housework support by affiliated enterprises.

## 4.3 Map-Oriented Contents

Since the above contents were created for a specific purpose, the information stopped after satisfying that purpose; reusing the information and cooperation between pieces of information were difficult. However, unitary management of this information was completed with the proposed model. After unification, suitable information for the present was displayed on the map, and information provided in such a form becomes a part of everyday life. Since the original information is the same, it can also become past contents whose details can be browsed on a map (Fig. 6).

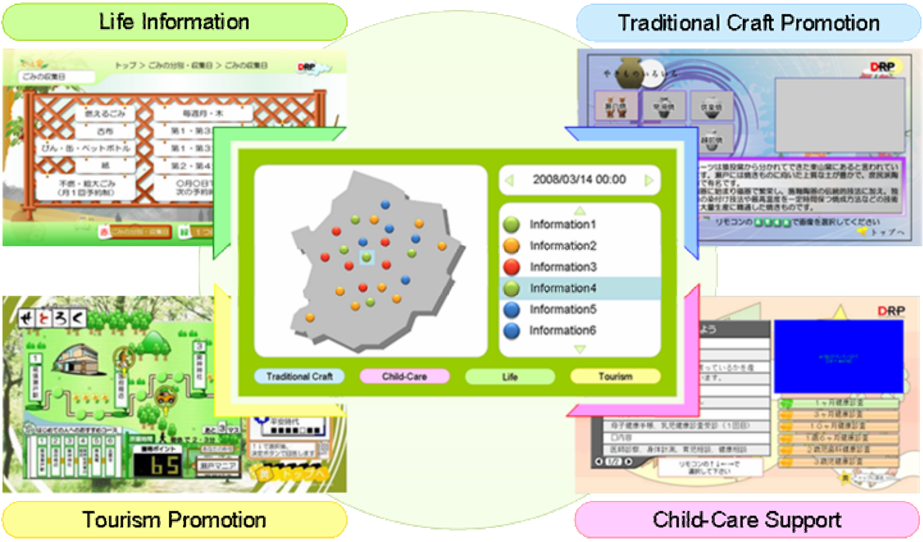


**Fig. 6.** Map interface

## 5 Conclusion

In this research, we proposed a map-oriented information model by focusing attention on a map as an information interface. Moreover, we showed applications to regional information on this model and created regional information contents for data broadcasting by information management based on it. This time, we only targeted a single region. When the information group formed by this model links areas, a sense of distance can be expressed in the real world between bits of information.

Now, various contents treat regional information including regional portal sites on the Internet. But they remain a single content in virtual space. However, when treated by this model, the trait of a "region" rises to the surface on the Internet. The creation of regional innovation based on regional information is promoted to increase the flexibility of the proposed model to have affinity with existing systems. We want to exhibit such specifications as defining by XML and supposing wide utilizations in society. Moreover, we want to practically apply our proposed model in various environments where local information is treated to verify its effects. We plan to research the ideal way for the Internet to use regional information mutually in online communities such as regional portal sites.

## Acknowledgments

This research was partially funded by Grants-in-Aid for Scientific Research, and High-Tech Research Center Project from the Ministry of Education, Culture, Sports, Science and Technology of Japan, and the Chukyo University Research Fund.

## References

1. Yamada, O.: Technical Trends of Digital Satellite and Terrestrial Broadcasting: Toward Second Generation Digital Broadcasting. IEICE technical report. Information theory 97(414), 1–10 (1997)
2. Ura, M., Harata, M., Niidome, Y., Hayashi, S., Nakamura, K., Yamada, M., Endo, M., Miyazaki, S.: Applying Digital Data Broadcasting to Community Informatization. Technical report of IEICE. Multimedia and virtual environment 105(566), 7–12 (2006)
3. Boiko, B.: Content Management Bible, 2nd edn. John Wiley & Sons Inc., Chichester (2004)
4. Association of Radio Industries and Businesses: Data Coding and Transmission Specifications for Digital Broadcasting, 4.0 (2004)

# Continuous Clustering of Moving Objects in Spatial Networks

Wenting Liu, Zhijian Wang, and Jun Feng

Hohai University, Nanjing, Jiangsu 210098 China
liuwt@hhu.edu.cn

**Abstract.** Spatial-Temporal clustering is one of the most important analysis tasks in spatial databases. Especially, in many real applications, real time data analysis such as clustering moving objects in spatial networks or traffic congestion prediction is more meaningful.Extensive method of clustering moving objects in Euclidean space is more complex and expensive. This paper proposes the scheme of clustering continuously moving objects, analyzes the fixed feature of the road network, proposes a notion of Virtual Clustering Unit (VCU) and improves on the existing algorithm. Performance analysis shows that the new scheme achieves high efficiency and accuracy for continuous clustering of moving objects in road networks.

**Keywords:** Clustering, Moving Objects, Road Networks.

## 1 Introduction

Clustering [1] denotes the grouping of a set of data items so that similar data items are in the same groups and different data items are in distinct groups. Clustering is widely studied in data mining and pattern cognition community. Clustering is very important to analysis the inner data structure. Early research mainly focused on clustering a static dataset [2,3,4,5]. With the increasing diffusion of wireless devices such as PDAs and mobile phones and the availability of geo-positioning, for example, GPS, a variety of location-based services are emerging. In recent years, clustering moving objects has been attracting increasing attention, which has various applications in the domains of traffic jam prediction, market research, medical figure auto-detection and weather forecast. However, little attention has been devoted to the clustering of moving objects and the most of existing work on clustering of moving objects assumed a free movement space and defined the similarity between objects by their Euclidean distance[6,7,8,9].

It is proposed that clustering moving objects is very meaningful in spatial networks in the literature [7]. In the real world, objects move within spatially constrained networks, e.g., vehicles in an urban move on road networks. Nowadays, there are a few studies on clustering nodes or objects in a spatial network and the studies mainly focus on the static dataset [7,8,9], and the studies on continuous clustering of moving objects in spatial networks is fewer [10], but the

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 543–550, 2008.

later is more meaningful, e.g., real-time monitoring the road running state, or traffic density region predication according to the current density region.

In the paper, we improve on the existing algorithm [10], and propose a new scheme for continuous clustering of moving objects at the intersection of road networks, not assuming the next edge.

Our contributions can be summarized as follows:

1) We propose a new data structure Virtual Cluster Unit (VCU) for efficient clustering moving objects at the intersection.
2) We develop a new group split scheme at the intersection of the road for reducing the processing cost and improving the processing accuracy.
3) We give the formal definition of VCU and the algorithms of creation and modification.

The rest of the paper is organized as follows. Section 2 surveys the related work. Section 3 details the scheme for continuous clustering of moving objects at the intersection of road networks. Section 4 shows performance analysis. We conclude this paper in Section 5.

## 2 Related Work

Many clustering techniques have been proposed for static data sets[1,2,3,4,5,11], the clustering approaches [6,8,12] of moving objects mainly based on Euclidean space. Yiu et al. first define the problem of clustering objects according to their network distance in SIGMOD 2004 [9] and propose algorithms that apply dominant clustering paradigms on the network-based clustering problem. The algorithms mainly process over spatial networks. The main focus of database research is how to organize large sparse networks on disk, such that shortest path queries can be efficiently processed, but it is inefficient to process dense networks. It focuses on the static objects that lie on spatial networks.Literature [10] proposes the way to clustering moving objects in the road networks, which extends the approach of literature [9]. Lai et al. extend the clustering approach in [10], propose the clustering moving objects in spatial network, define the cluster unit (CU) formally. A CU [10] is a group of moving objects close to each other at present and near future time. The objects in a CU move in the same direction and on the same segment. The CUs' maintenance includes two phases: one is the phase moving on the segment; the other is the phase arriving at the end of the segment. On the segments the main task is to dynamically maintain the order of objects and compute the valid time. At the end of segments a group split scheme is proposed according to the objects' next segment, to reduce the processing cost. In real application, the group split scheme can't be realized, as on one side we can't get the next segment, on the other side when a few CUs on different segments arrive at the same intersection, we must perform parallel group scheme, compute the valid time of each CU and maintain the interlaced orders. So the scheme is very complex.

In order to handle above questions, we propose a new clustering scheme in the intersection, not assuming to known the next edge to move along.

# 3 Continuous Clustering of Moving Objects in the Intersection of Roads

## 3.1 Modeling of Moving Objects

We model a road network as a graph [9]. Nodes of the graph represent intersections of road network and edges represent segments. Objects are moving on the edges. Any object can only locate on a segment. The distance between any two objects, called network distance, is measured by the length of the shortest path connecting them in the network. We employ a similar motion model as in [9], where each object is assumed to move at a stable velocity at each edge. Each object is capable of transmitting its current location and velocity to a central server. Each object location update has the following form $(o_{id},n_a,n_b,pos,v)$ where $o_{id}$ is the id of the moving object, $(n_a\ ,n_b)$ represents the edge on which the object moves from $n_a$ toward $n_b$ , $pos$ is the relative location to $n_a$, and $v$ is the moving speed. The clustering processing of the existing algorithm [9] includes two phases:

1) A set of clustering unit are created by traversing all segments in the network and their associated objects. The CUs are incrementally maintained after their creation.
2) As time elapsed, the similarity between adjacent objects in a CU or between CUs may change, so CUs need be maintained dynamically. Especially, when CUs arrive at the intersection of road networks, we need perform CUs' split. In order to reduce the split and merge cost and improve the accuracy, we propose the scheme based on virtual cluster units (VCU).

## 3.2 Virtual Cluster Unit (VCU)

Formally, a VCU is defined as follows:

**[Definition 1]** *Virtual Cluster Unit (VCU)***.** *A VCU is represented by VCU=(* $VCU_{id}, O, n_a, n_b, head$*), where* $VCU_{id}$ *is the id of the VCU, O is a list of objects* $\{o_1, o_2, \ldots, o_i \ldots o_n\}$, $o_i = (o_{id_i}, n_a, n_b, pos_i, v_i)$*, where* $pos_i$ *is the relative location to* $n_a, v_i$ *is the moving speed. Without loss of generality, assuming* $pos_1 \leq pos_2 \leq \ldots \leq pos_n$*, Since all objects locate at the beginning of the segment* $(n_a, n_b)$*, of which direction is from* $n_a$ *to* $n_b$*, the position of the VCU is determined by an interval (*$n_a$*, head) in terms of the network distance from* $n_a$*. Thus the length of the VCU is* $|head - n_a|$*.* □

VCU is formed by split and merge of CUs. So VCU is closely connect with CU. The following definition is the existing definition in [10].

**[Definition 2]** *Cluster Unit (CU)***.** *A CU is represented by CU= (*$CU_{id}$ *, O ,* $n_a$ *,* $n_b$*, head, tail), where* $CU_{id}$ *is the id of the CU, O is a list of objects*$\{o_1,o_2,\ldots,o_i$ $\ldots, o_n\ \}$, $o_i = (o_{id_i}, n_a\ , n_b, pos_i, v_i)$*, where* $pos_i$ *is the relative location to* $n_a$*,* $v_i$ *is the moving speed. Without loss of generality, assuming* $pos_1 \leq pos_2 \leq \ldots \leq pos_n, pos_{i+1} - pos_i \leq \epsilon(1 \leq i \leq n-1)$*, Since all objects are on the same segment* $(n_a, n_b)$*, of which direction is from* $n_a$ *to* $n_b$ *,the position of the CU* *is determined by an interval (head, tail) in terms of the network distance from* $n_a$*. Thus the length of the CU is* $|tail - head|$. □

The forms of CU and VCU are similar. Both are a group of moving objects. However, CU is a micro-cluster [7], the objects in a CU are similar and the objects in different CU are dissimilar. CUs can locate on segments. VCUs can only locate at the beginning of segments. There are only one VCU on a segment and the objects in a VCU may be dissimilar.

### 3.3 Clustering Scheme in the Intersection of Roads

**Clustering scheme.** When CUs arrive at the end of the segment, the processing would be more complex. The number of CU splitting is connected with the number of adjacent segments. To reduce the processing cost, we propose a new group split scheme. The object transmits its current location and velocity to a central server when the object moves to the adjacent segment. When the $CU_i$ reaches the end of segment $node_i$, we perform the split event. We address the time as $t_{is}$ when the first object of $CU_i$ leaves, and we compute the time $t_{ie}$ of the last object leaving. So the valid time of the CU is $[t_{is}, t_{ie}]$. During the valid time, if there is another $CU_j$ reaches the intersection, we perform parallel group scheme, and compute the valid time $[t_{js}, t_{je}](t_{is} < t_{js} < t_{ie})$. So the valid time of the intersection $node_i$ is $[t_{is}, \max(t_{ie}, t_{je})]$. Fig.1 illustrates the procession of valid-time of VCU. During the valid time, we add each leaving object to VCU, if exist VCU in the adjacent segment. Otherwise, we firstly create a new VCU in the adjacent segment. When the time is expired, we directly delete all the CU of the intersection and append all the objects of VCU to the tail of adjacent CU (the distance of CU and VCU is not greater than $\epsilon$). At last we delete the VCUs.

Fig.2(a) illustrates the procession of VCU's creation. When $CU_1$ reaches the end of the segment $n_e$, we compute $CU_1$'s valid time $[t_1, t_2]$. During the valid time, objects $o_2$ and $o_4$ move to segment $(n_e, n_b)$, $o_1$ and $o_3$ move to segment $(n_e\ , n_c)$. When $CU_2$ reached the node $n_e$, we compute $CU_2$'s valid time $[t_3, t_4]$ (i.e., $t_1 < t_3 < t_2 < t_4$). Objects $o_5$ and $o_7$ move to segment $(n_e, n_b)$, $o_6$ and $o_8$ move to segment $(n_e, n_c)$. At the time $t_2$, the configuration is formed (i.e., Fig.2(b) parallel splitting of CUs). We don't delete $CU_1$ and $CU_2$ until the last

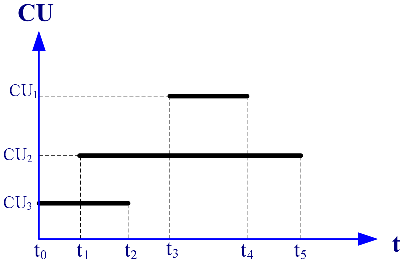


**Fig. 1.** Valid-time of VCU

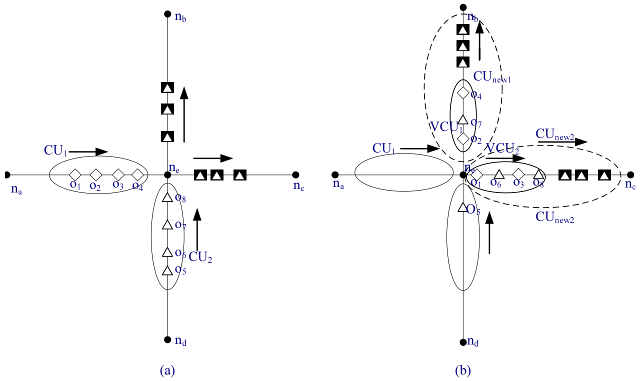


**Fig. 2.** (a)Creation of Virtual CU. (b)Parallel splitting of CUs.

object $o_5$ arrives at the segment $(n_e\ , n_b)$ and perform the VCU created. The new group split scheme at the intersection of the road keep the effective of parallel splitting, maintain the reasonable objects order of the parallel CUs and reduce the splitting times so that reduce the maintenance cost of CUs.

**The Lifetime of VCU.** The lifetime of VCU includes 3 phases: creation, running and destroy.

1) Creation: an object o moves to the adjacent segment $(n_a,\ n_b)$, where there isn't a VCU, we create a new VCU.
2) Running: when an object o arrives at the intersection, we directly add the object o to the VCU.
3) Destroy: when the valid time of VCU is over, we merge the objects of VCU to the adjacent CU and delete VCU directly from queue.

**The Modification of VCU: Insertion and Deletion.** We firstly compute the valid time of VCU when VCU is modified. A few CUs may arrive at the intersection of the segments at the same time. During the valid time, we don't change the state of the parallel CUs and add the leaving objects to the VCUs. When the objects leave to the next segment, update the objects information $(o_{id}, n_a, n_b, pos_i, v_i)$ and add the object to the order list of VCU, but don't split and merge the CU and VCU. The Algorithm 1 illustrates the insert process.

=====================================

**Algorithm 1. Insert(o)**

```
/* Input: o is an object to be inserted*/
begin
  for each $node_i$ do
    if $node_i$ .visited==false then
      Q=new priority queue;
      $VCU_{id}$ ←find the VCU of the edge where o lies
      if $VCU_{id}$==null then
```

```
7.              create a new $VCU_{id}$ ;
8.           if o is not the first object of $CU_i$ then
9.              Enqueue(Q,$CU_i$);
10.             insert o into objects list of $VCU_id$
     end Insert
=========================================
```

We process all the objects of the CUs, respectively. When the first object of the parallel CUs leaves to the adjacent segment, we create a new VCU (lines 3 to 7). On the contrary, when the leaving object is the first object, we add the object to the VCU directly (lines 8 to 10).

When VCU or CU is not in valid time, we delete all the CUs of the intersection of segments, delete all the split event of the CUs and process the split and merge event of VCU. We search the first CU where VCU lies. If we can search the CU which the distance of CU and VCU is not greater than , we address it as FirstCU, otherwise, we create a new CU, addressed newCU. Then merge VCU to FirstCU or newCU, at last process the split and merge event of the merged CU. The Algorithm 2 illustrates the procession of valid-time of VCU.

```
=========================================
```

**Algorithm 2. Delete()**

```
   begin
1.       for each $node_i$ do
2.          while notempty(Q) do
3.          CU=Dequeue(Q);
4.          delete CU ;
5.          delete all CUs' event from the event queue;
6.       for each $VCU_i$ of $node_i$ ;
7.          firstCU ← find first CU on the same edge with VCU
8.          if firstCU $\neq$ null then
9.             minDis←compute the minimize distance between $VCU_i$ and firstCU
10.         if minDis $> \epsilon$ or firstCU==null then
11.            create a new CU as firstCU
12.         Expand_Merge($VCU_i$, firstCU)
13.         delete $VCU_i$ from $VCU_i$ queue;
   end delete
=========================================
```

We process all the CUs and VCUs, respectively. First we delete all the CU in the priority queue Q (lines 3 and 4) and delete all CUs' event from the event queue(lines 5). Then we process the VCUs. If we can find the appropriate the adjacent CU of VCU (lines 7 to 9), otherwise, we create a new null CU (lines10 and 11). At last merge the VCU to the next CU and delete VCU from VCU queue (lines 12 and 13).The following procedure Expand_Merge merge VCU to the adjacent CU.

```
==============================================
Procedure Expand_Merge(VCUfirstCU)
  begin
        Olist← get the O list from VCU
        insert Olist to the tail of firstCU
          split(fisrtCU);
  end Expand_Merge
==============================================
```

# 4 Performance Analysis

In this section, we evaluate the performance of our proposed techniques by accuracy and cost analysis. The analysis shows our algorithm is accurate and efficient.

## 4.1 Accuracy

In our approach, we mainly process the CUs' split at the intersection. The processing includes a single CU's split and mutil-CU's split. For the single CU's split the accuracy is same with the existing algorithms [10]. For the mutil-CU's split, the accuracy is higher than the existing algorithms, since using parallel splitting scheme.

## 4.2 Cost

Assume there are N moving objects at the intersection and M CUs ($CU_1$, $CU_2$ $\ldots$ $CU_M$) that contain $N_1$, $N_2$ $\ldots$ $N_M$ moving objects respectively. There are E valid edges and V nodes in the road network. Usually $M \ll N$ or M is 2E at most and $\sum N_i$=N. Let $|Q|$ be the length of event priority queue Q, which has a size of O(SQ+MQ), where SQ and MQ are the number of split and merge events stored in Q, respectively.

In our approach, we only take into account the CUs at the intersection. Computing a split event from a CU takes $M \log(SQ + MQ) \sum N_i$ time ,which is O(M.SQ.log($N$) $\log(SQ + MQ)$) and inserting a split event to the priority queue Q takes $O(\log(|Q|))$. The cost of computing all the initial split events is O(Mlog($|Q|$)+ M.SQ.$\log N$log($SQ$+$MQ$)). The maintenance phase of VCUs processes insertion/deletion event. It requires $O(M + |E| \log(|E|))$ time to build the initial VCUs at the intersection. The insertion events require $O(|E| \log(|E|))$. Deletion events require $O(M \log(|Q|) + N^2/M + |E|^2)$. The total cost of the phase is $O(Mlog|Q| + N^2/M + |E|^2 + |E| \log(|E|))$ .For our approach only refers to the intersection, the cost of the approach is O(M.SQ.log N log(SQ + MQ)+Mlog $|Q|$+ $N^2$/M +$|E|^2$+$|E| \log |E|$).

According to[10], the cost of the group scheme at the intersection is O(M. SQ. $\log N$log(SQ + MQ)+Mlog $|Q|$+ $N^2/M$ +$|E| \log |E|$). The cost of[10] at the intersection only reduces constant time $|E|^2$, but our accuracy is higher. Thus our approach is efficient.

## 5 Conclusion

In this paper, we proposed a new group split scheme for clustering moving objects at the intersection, reduced the processing cost and improve the accuracy of the existing algorithm based on the intersection of road. The performance Analysis showed the efficiency of our method. In the future, we plan to further investigate the applicability of this method. We will also evaluate the performance by experiment based on a real road network.

## Acknowledgements

This research is supported by NSFC 60673141 (Research on the Index Structure of Spatial Network-based Moving Objects).

## References

1. Jain, A.K., Dubes, R.C.: Algorithms for Clustering Data. New Jersey Prentice-Hall Advanced Reference Series, pp. 1–334 (1988)
2. Agrawal, R., Gehrke, J., Gunopulos, D., Raghavan, P.: Automatic Subspace Clustering of High Dimensional Data for Data Mining Application. In: Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD 1998), pp. 94–105 (1998)
3. Ankerst, M., Breunig, M., Kriegel, H.P., Sander, J.: OPTICS: Ordering Points to Identify the Clustering Structure. In: Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD 1999), pp. 49–60 (1999)
4. Ng, R., Han, J.: Efficient and Effective Clustering Method for Spatial Data Mining. In: Proc. 20th Int'l Conf. Very Large Data Bases (VLDB 1994), pp. 144–155 (1994)
5. Kaufman, Rousseeuw, P.J.: Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley and Sons, Inc., Chichester (1990)
6. Kalnis, P., Mamoulis, N., Bakiras, S.: On Discovering Moving Clusters in Spatiotemporal Data. In: Bauzer Medeiros, C., Egenhofer, M.J., Bertino, E. (eds.) SSTD 2005. LNCS, vol. 3633, pp. 364–381. Springer, Heidelberg (2005)
7. Li, Y., Han, J., Yang, J.: Clustering Moving Objects. In: Proc. 10th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining(KDD 2004), pp. 617–622 (2004)
8. Tung, A.K.H., Hou, J., Han, J.W.: Spatial clustering in the presence of obstacles. In: Proc. of the 17th Int'l Conf. on Data Engineering (ICDE), pp. 359–367. IEEE Computer Society, Heidelberg (2001)
9. Yiu, M.L., Mamoulis, N.: Clustering Objects on a Spatial Network. In: SIGMOD, pp. 443–454 (2004)
10. Lai, C., Wang, L., Chen, J., Meng, X., Xu, J., Zeitouni, K.: Effective Density Queries for Moving Objects in Road Networks. In: Dong, G., Lin, X., Wang, W., Yang, Y., Yu, J.X. (eds.) APWeb/WAIM 2007. LNCS, vol. 4505, pp. 200–211. Springer, Heidelberg (2007)
11. Li, Y.J.: A clustering algorithm based on maximal -distant subtrees. Pattern Recognition 40(5), 1425–1431 (2007)
12. Jensen, C.S., Lin, D., Ooi, B.C.: Query and Update Efficient B-Tree Based Indexing of Moving Objects. In: 30th Int'l Conf.Very Large Data Bases (VLDB 2004), pp. 768–779 (2004)

# Index Method for Tracking Network-Constrained Moving Objects

Jun Feng[1], Jiamin Lu[1], Yuelong Zhu[1], and Toyohide Watanabe[2]

[1] Hohai University, Nanjing, Jiangsu 210098 China
{fengjun,welmanwenzi,ylzhu}@hhu.edu.cn
[2] Nagoya University, Nagoya, Aichi 464-8603 Japan
watanabe@is.nagoya-u.ac.jp

**Abstract.** In the past composite structures for managing road network-constrained moving objects, the road network is usually broken up into road segments. This scheme will cause high concentrated update operations, and some needless queries while tracking the moving objects. In this paper, we propose a new unit called cross region(CR) to break up the road network, and use CR to build a new structure called CR-tree to be the static part of the composite structure. By indexing the moving objects with our composite structure, experiments show that the update density could be evened, the update frequency and the update cost could be decreased compare with the past ones.

## 1 Introduction

In recent years, with the advancement of portable computing, geographic positioning and wireless communicating technologies, the devices like PDA, GPS become more cheaper, location-based service (LBS) in the urban traffic system, e.g., traffic navigation, traffic flow forecasting becomes possible. In such kinds of systems, tracking the vehicles real-timely plays an important role, which needs an efficient method for collecting and managing the network-constrained moving objects' positions.

To manage the moving objects, there are a lot of structures like TPR-tree[1], TPR*-tree[2], 3DR-tree[3] and the SV model[4] proposed in recent years. These structures manage the objects moving in the Euclidean Space, where objects can move to any place in a straight line, e.g., planes move in air. However, in the urban transportation that this paper mainly concerns, objects' movements are constrained, not only their moving directions but also the distance among them are constrained by the road network.The neighboring relations among these moving objects should take the underlying road network into consideration.

For managing the *road network-constrained moving objects*, many composite structures[5] were proposed, e.g., NCO-tree[5], FNR-tree[6], IMTFN[7], MON-tree[8] and MONC-tree[9]. In these structures, road networks are broken up into different units, usually road segments, and indexed by a spatial R-tree-like structure. The moving objects inside every unit are indexed by specific spatiotemporal structures. For tracking the moving objects, when vehicles run from one unit to

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 551–558, 2008.

another, their position data should be inserted into the correct units' dynamic structures. In the situation of regarding road segment as the unit, there are two problems in the this process: uneven update density and high update cost. For example, in Fig.1, road network ***RN*** with nodes $\{a, b, c, d, e, f, g, h\}$ is broken up into *segments* $\{ab, ac, ad, ae, bd, ef, fg, fh, gh\}$. Assume there is *Car* in *ae* with speed $\mathbb{v}$, it will move to *ab* after passing through *a*. Here, when *Car* passes through *a* after sometime later, it should submit an update request to the system, then *Car*'s data will be moved from *ae*'s spatiotemporal structure to *ab*'s in time. If this update is not quickly enough, the accurate position of *Car* will be lost in the index. Therefore, most vehicles' updates are concentrated around the crosses, and make the update density uneven. At the same time, as there is not any connection information about segments (e.g., the connection information about *ab* and *ae*) inside the index, when process the update operation, the system should find out *ab* through *Car*'s new position. Especially, in day time, not only a number of vehicles run across a crossroad in a short time[10], but also the number of crosses in the urban road network are all very large, the whole update cost of the composite structure will be high.

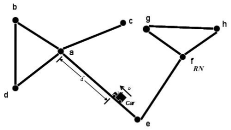


**Fig. 1.** Example of Road Network ***RN***

In this paper, we solve the up problems by proposing a new index structure for managing road network: CR-tree. *Cross region*(CR) is our basic unit to break up the road network. A CR covers a region around a cross in road network. Two CRs are adjacent if the two crosses they covered respectively are adjacent too, and this adjacent relationship will be kept in both CRs. When a composite structure adopts CR-tree for managing the road network, we can get three benefits:

– **Update Cost.** Because the adjacent relationships are kept in CRs, systems can "know" the CRs part vehicles will enter. Therefore, the update cost could be decreased.
– **Update Density.** The update operations could be dispersed into any place inside a CR, not concentrated around the cross, so that the update density of the structure could be evened.
– **Update Latency.** When we use CR to break up the road network, part vehicles' update latency could be extended further, so that the total amount of updates in a same time slot could be decreased.

In the following of this paper, Section 2 gives the definition of CR, and analyze the benefits brought by CR. Section 3 depicts an improved R*-tree-like method, CR-tree, used to index the CRs, section 4 gives the analysis with experiments, and conclusion is made in section 5.

## 2 Cross Region

This section will introduce the definition of the Cross Region(CR), and analyze its effect on update density and update latency.

### 2.1 The Definition of CR

A CR is defined as a quad-tuple: ($cID$,*Cross*-*Array*, $MBR$, *OR*-*Array*). Here, $cID$ refers to the ID of the cross it contains, we use $CR_i$ to denote the CR for cross $i$. *Cross*-*Array* contains IDs of the neighboring crosses, and $MBR$ is the minimum bounding rectangle of this CR in 2D space. As CR will cover part of every segment it attached with, the overlap ratios will be recorded in *OR*-*Array*. E.g., in Fig.2, $CR_a$ contains the cross $a$, *Cross*-*Array* is $\{b, c, d, e\}$, and the dashed rectangle is its MBR. Those solid lines inside the MBR means the roads covered by $CR_a$, and the decimals beside the solid roads is the ratio of the road to the segment it belongs to, these ratios are all decimals in $(0, 1)$, and $CR_a$'s *OR*-*Array* is $\{0.5, 0.8, 0.4, 0.3\}$.

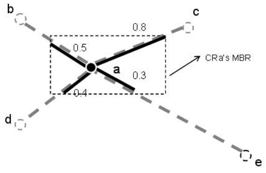


**Fig. 2.** Example of Cross Region cover the cross $a$ in ***RN***

### 2.2 Update Density with CR

In this section, we will analyze the cross region with example, to explain how CR could even the update density of the composite structure.

In Fig.3, there are two CRs: $CR_a$ and $CR_b$. The vehicle *Car* is moving on the segment $ab$ from $a$ to $b$ with speed $v$. The set $CS_a$ contains $m$ crosses abut with $a$, and the set $CS_b$ contains $n$ crosses abut with $b$. *Car* should submit its update at one of the situations:

- **Situation ①:** *Car* is leaving a cross and going to move out of the CR. As *Car* is leaving cross $a$ and moving along the segment $ab$ inside the scope of $CR_a$, it will move into the scope of $CR_b$ after less than $\delta\mathfrak{t} = \frac{L(ab)*CR_a.OR_b}{v}$ time later with high possibility. Where the function $L(i)$ means the length of segment $i$, and the $CR_i.OR_j$ is the overlap ratio of the road inside $CR_i$ to the segment between cross $i$ and $j$.
- **Situation ②:** *Car* runs into a new CR. When *Car* runs into $CR_b$, it will run inside the scope of $CR_b$ at most $\delta\mathfrak{t} = \frac{L(ab)*CR_b.OR_a+min\{L(bi)*CR_b.OR_i|i\in CS_b\}}{v}$.

The update at any situation can "forecast" that: *Car* will run into a new CR after $\delta\mathfrak{t}$ time later if there is no more update submitted from *Car*. It means although vehicles still have to update once when they are inside a CR, the

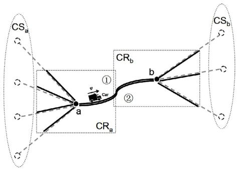


**Fig. 3.** The analysis of update density with CR

update requests could be submitted at any place of the road segments, not have to concentrate around the crosses, so that the update density of the structure could be evened.

### 2.3 Update Latency with CR

We denote the scheme to break up the road network into segments and crosses as *Segment Scheme*, and use *CR Scheme* to name the scheme of breaking up the road network into CRs.

In *Segment Scheme*, vehicles usually submit their update requests at Situation ① as we mentioned in last section, and the longest update latency of *Segment Scheme* is: $t = \frac{L(ab)}{v}$ But in *CR Scheme*, the longest update latency $t'$ at Situation ① is: $t' = \frac{L(ab)+min\{L(bi)*CR_b.OR_i|i\in CS_b\}}{v}$

Situation ② will also happen in *Segment Scheme* sometimes, e.g., vehicles may be started in the middle of the segments. The longest update latency $t$ with *Segment Scheme* is: $t = \frac{L(ab))*CR_b.OR_a}{v}$ But with *CR Scheme*, the longest update latency $t'$ is: $t'' = \frac{L(ab)*CR_b.OR_a+min\{L(bi)*CR_b.OR_i|i\in CS_b\}}{v}$

Through the comparison, the update latency $t$ of *Segment Scheme* is always shorter than $t'$ of *CR Scheme* at both situations, which means the update latency is extended in our method.

## 3 Structure of CR-Tree

In this section, we will use an improved R-tree-like method to index the CRs, called CR-tree. With CR-tree, particular CR could be searched via its position, and its adjacent CRs could also be found through the connection relationships among them.

### 3.1 Leaf Node

Follow the index method which used in R*-tree, Cr-tree will also keep the data of CRs in its leaf nodes. The CRs which are close by each other will be stored in the same leaf node. But different from the segment method, the CRs in the same leaf node could make up a connected sub-road-network, which means vehicles could move from any CR to the others in the same leaf node. The form of the leaf node is: $(nID, entry-Array, adjacent-matrix, MBR)$ Here, $nID$ is the unique identifier of the node, *entry-Array* is the array of inside CR entries, and $MBR$ is the minimum bounding rectangle this node overlaps in 2D space. The entries belonging to *entry-Array* are called as *"Inside Entry"*. The entries for adjacent CRs are called as *"External Entry"*. The *adjacent-matrix* will record not only the adjacent relationships among the *"Inside Entries"*s, but also the relationships among the *"Inside Entries"*s and *"External Entries"*.

## 3.2 Internal Node

The form of internal node is the same as leaf node, and its *"Inside Entries"* could also compose a connected sub-road network, except that the entries in its *entry-Array* doesn't point to CRs, but point to a sub-node in the CR-tree. When there is at least one pair of adjacent CRs belong to different sub-nodes, this two entries are connected in internal node. As there may be exist more than one pair of adjacent CRs, so we only use 'True' or 'False' to show two sub-nodes are connected or not. The adjacent relationships among the *"Inside Entries"* and *"External Entries"* are also recorded in internal node's *adjacent-matrix*.

## 3.3 Register Table of CR

For achieving the leaf node contains the target CR quickly, we use a "Register Table" for recording the addresses of CRs in the leaf nodes. Once a new CR is inserted into the CR-tree, the address of the new CR is recorded into the "Register Table". If a leaf node is combined or split, the "Register Table" will be updated.

Fig.4 shows the road network and its corresponding CR-tree. Here we only use $i$ to denote the $CR_i$ which contains the cross $i$ for simplicity. Each CR has its offset of the "Register Table" shown at the right of the CR-tree, and the table contains the addresses of the leaf nodes. If a vehicle updates in $CR_a$, and its update position is at situation ① on segment $ae$, via CR-tree, we will find it is inside $CR_a$ and is going to another CR which offset in "Register Table" is 4. Then we will achieve the leaf node $B$ according to the address recorded in the table at the offset 4. After comparing the offsets of the CRs $B$ contains, we can find $CR_e$ is the next CR this vehicle will move to.

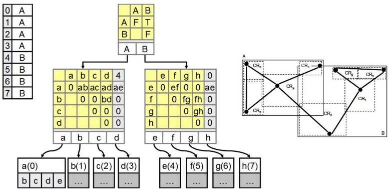


**Fig. 4.** The structure of Cr-tree

### 3.4 Operations of CR-Tree

The search algorithm of the CR-tree is to find out the leaf node contains the target CR, the procedure of the search operation is:

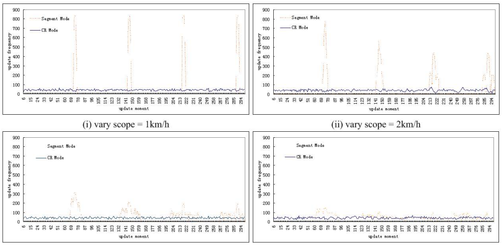
```
SEARCH CR_Search(T, tcr, *ln)
/*Input: T is root node of a Cr-tree, and tcr is the target CR,
Output: whether T contains tcr, if find, return the node pointed by ln*/
if (T.level $\neq$ 0) /*T is a internal node*/
  for each subNode in T.sn-Array
    if (subNode.MBR overlap tcr.MBR)
      if CR_Search(subNode, tcr, ln)
        return True
else /*T is a leaf node*/
  for each cr in T.cr-Array
    if tcr.ID == cr.ID
      ln = &node
      return True
```

The process of inserting a new CR into Cr-tree is similar to the insertion in the R*-tree, the difference is that the new CR should be adjacent with the leaf node it will be inserted into. Nodes that overflow are split. The **Insert** will use the **ChooseLeaf** algorithm, which is used to select a leaf node that the ratio of the node's MBR's diagonal to the segment between the new CR and the node is minimum to place the new CR.

The split algorithm is used to divide an overfull node $N$ into two nodes $N$ and $NN$. Similar with traditional spatial structures, the method on dividing MBR is the same as R*-tree. The difference is that, as the *"Inside Entries"* in $N$ could compose a connected sub road network, therefore, if the *"Inside Entries"* belong to $N$ and $NN$ respectively, after splitting, they must also could compose two small connected sub road network. For this purpose, we will divide the *entry-Array* simply into two different entry sets of as near equal sizes as possible, then check each set to find out those entries not connected with other entries in the same set, and put these unconnected entries to another set. We omit these procedures for the paper length limitation.

## 4 Analysis

In this section, we simulated the movements of vehicles running on the road network: about 3000 vehicles move ten kilometers with similar speeds, pass through same crosses, and update with same mode. These vehicles use two update modes when they run through these roads, one mode is segment mode, which means we break up the road network into segments, and vehicles update their positions once they enter into a new segment; the other mode is CR mode, which means we break up the road network into CRs, and vehicles submit updates while they



(i) vary scope = 1km/h

(ii) vary scope = 2km/h

(iii) vary scope = 5km/h

(iv) vary scope = 10km/h

**Fig. 5.** The comparison on the update density

are inside each CR. As we discussed before, there are two situations when we use the CR mode, so in the simulation, about half vehicles will always update in Situation ①, and the other vehicles will always update in Situation ②.

For analyzing the update density under different update mode, we count the update times at each moment when vehicles run through the road network with same speeds, and the results are showed in Fig.5.

As the ideal traffic situation is all the vehicles move with the same speed, so in this simulation, we set the average speed of the vehicles take part in, but the specific velocity of each vehicle will plus or minus an random value inside a scope. In the simulation shown in Fig.5, the average speed of these vehicles is 50km/h, and every experiment has a different vary scope. In Fig.5, the dash lines show the update density with segment mode and the solid lines show the update density with CR mode.

In Fig.5(i), as the scope of the vehicles' velocities is 1km/h, which means their velocities are random number inside $[49, 51]$km/h. In this situation, most updates with segment mode are constrained around several moments, because vehicles all run through the crosses at these moments. But when these vehicles use CR mode to submit their updates, the update density become evener. The Fig.5(ii)'s scope is 2km/h, in this situation, the update density with segment mode become evener than Fig. 5(i), but it still unevener than the situation with CR mode. The results of setting scope as 5km/h and 10km/h which are shown in Fig.5(iii) and Fig.5(iv) have the same tendency.

From the experiment results, we can find out the update density with segment mode is effected by the scope these vehicles' velocities vary inside. When the scope is wider, the update density will be evener. But the update densities with CR mode keep steady with different vary scope, and all have evener curves than the segment mode.

## 5 Conclusion

In this paper, aim to the management of road network-constrained spatiotemporal data, we improve the static index part of the composite structure which is usually used in past research. We propose a new unit called Cross Region(CR) instead of segment to break up the road network, and index the CRs into a

CR-tree to be the static part of the composite structure. With CR, the update latency of the whole structure could be extended because some vehicles' movements could be anticipated longer than with segment in some conditions, which could be proved theoretically. Secondly, as vehicles could submit their updates wherever inside the segments, so the update density could evened when we use CR instead of segment, which could be proved by experiments. Thirdly, we designed the register table of CRs, each CR could find the pointer to the leaf node contains it through its unique offset inside this table, so that vehicles' data could be move to the adjacent CR from the present one, and the update cost of the composite structure could be decreased.

## Acknowledgements

This research is supported by NSFC 60673141 (Research on the Index Structure of Spatial Network-based Moving Objects).

## References

1. Civilis, A., Jensen, C.S., Pakalnis, S.: Techniques for Efficient Road-Network-Based Tracking of Moving Objects. IEEE Trans. Knowl. Data Eng. (TKDE) 17(5), 698–712 (2005)
2. Tao, Y., Papadias, D., Sun, J.: The TPR*-Tree: An Optimized Spatio-Temporal Access Method for Predictive Queries. In: VLDB 2003, pp. 790–801 (2003)
3. Theodoridis, Y., Vazirgiannis, M., Sellis, T.K.: Spatio-Temporal Indexing for Large Multimedia Applications. In: ICMCS 1996, pp. 441–448 (1996)
4. Chon, H.D., Agrawal, D., Abbadi, A.E.: Stoage and Retrieval of Moving Objects. Mobile Data Management, pp. 173–184 (2001)
5. Feng, J., Lu, J., Zhu, Y., Mukai, N.: Toyohide Watanabe: Indexing of Moving Objects on Road Network Using Composite Structure. In: KES 2007, pp. 1097–1104 (2007)
6. Frentzos, E.: Indexing Objects Moving on Fixed Networks. In: SSTD 2003, pp. 289–305 (2003)
7. Li, G., Zhong, X.: Indexing Moving Objects Trajectories on Fixed Networks. Journal of Computer Research and Development, 828–833 (2006)
8. de Almeida, V.T., Gting, R.H.: Indexing the Trajectories of Moving Objects in Networks. GeoInformatica 9(1), 33–60 (2005)
9. Li, X., Fu, P., Liu, J.: Spatio-temporal Index Based on Networks. Geomatics and Information Science of Wuhan University 31(7), 620–623 (2006)
10. Zhu, H., Zhu, Y., Li, M., Ni, L.M.: ANTS: Efficient Vehicle Locating Based on Ant Search in ShanghaiGrid. In: ICPP 2007, p. 34 (2007)

# Adaptive Routing of Cruising Taxis by Mutual Exchange of Pathways

Kosuke Yamamoto, Kentaro Uesugi, and Toyohide Watanabe

Department of Systems and Social Informatics
Graduate School of Information Science, Nagoya University
Furo-cho, Chikusa-ku, Nagoya, 464-8603, Japan
{yamamoto,uesugi,watanabe}@watanabe.ss.is.nagoya-u.ac.jp

**Abstract.** The tremendous development of information and communication technology had large influence for service management in the taxi dispatching work. However, taxi drivers who want to drive in "cruising taxis" decide their travel routes by depending on their own heuristics. As a result, traffic jams and local excess supplies have often been occurred. In this paper, we propose an adaptive routing method in the cruising taxis. In our method, pathways where many customers are expected to exist are assigned to drivers. This assignment changes dynamically adapting to changes of taxis' positions. Our simulation experiment shows that our method was able to gain more customers than the existing means of cruising taxis.

## 1 Introduction

Taxi is an important transport means which can carry customers by door-to-door. Over a few past years, several studies have been reported on taxi services with information and communication technology. In particular, the taxi dispatching system which uses Global Positioning Systems (GPS) [1] came into practical use and became widely prevalent. In urban areas, a number of taxis run around here and there so as to pick up timely some customers on the pathways. This type of taxis is called "cruising taxi". Generally, cruising taxis take the routes which are established by their drivers individually. This makes taxis saturated at areas where many customers are thought possibly to exist. As a result, empty taxis, which are not carrying any customers, cause traffic jams. Moreover, operating revenues of taxi companies decrease.

We focus on the routing problem of cruising taxis and aim to compute the travel routes so that cruising taxis can gain more customers. We denominate this problem "Multiple Traveling Taxi Problem(MTTP)". MTTP is similar to Multiple Traveling Salesman Problem(MTSP)[2] in that multiple vehicles travel routes in the area. One approach to solve MTSP is to make use of the clustering algorithm. However, static clustering algorithms are not effective for MTTP because demands of customers change positional distribution of taxis. Thus, we utilize the concept of fuzzy clustering [3]. The fuzzy clustering is one of clustering algorithms that vaguely divide objects into clusters by membership values. This

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 559–566, 2008.

fuzzy clustering is suitable to MTTP because it is flexible to changing of elements in clusters. In the fuzzy c-means clustering [4], which is the famous algorithm in the fuzzy clustering, the membership value is computed on the basis of Euclidean distance between an object and the representative point of a cluster. We define the membership value which is suitable to MTTP and propose the method of dynamic reassembly of clusters by mutual exchange of pathways.

The remainder of this paper is organized as follows. In Section 2, we refer to our approach. Section 3 formulates MTTP. In Section 4, we mention an assignment method of taxis' travel routes. In Section 5, we report simulation results and speculate about effectiveness of our method. Section 6 concludes this paper and offers our future work.

## 2 Approach

The issue which cruising taxis face to today is that it is not always easy for taxi drivers to gain many customers because of their selfish decisions of traveling routes. In order to reduce the locally excessive supplement of taxis, the cruising taxis must travel concertedly. It is difficult for each driver to assess the situation evermore and cooperate with each other effectively. The method to maintain the efficient assignment of routes based on the appearance frequency of customers is requisite. In this paper, we assume that taxi drivers can infer this appearance frequency accurately from their daily logs.

Most customers which cruising taxis anticipate appear on the road: for this reason, MTTP can be set down as a particular kind of MTSP by substituting some roads for cities in MTSP. The feature of MTTP is that when taxi drivers gain customers they must move to customers' destinations and taxi drivers cannot estimate the destinations accurately. In addition, since it is preferred that travel routes of taxis contain more proportions of pathways where many customers exist, clustering methods based on representative points, such as fuzzy c-means clustering, are not appropriate to MTTP. In order to increase the total benefit, a pathway should be passed by the taxi which can come through the pathway easier than any other taxis. Therefore, for adaptive assignment of routes, we define the membership value based on the cost to pass the pathway during traveling and utilize it.

In our method, taxis exchange mutually pathways in their travel routes based on the membership value. The method enables taxis to gain more customers effectively.

## 3 Formalization

A road network in a travel area is given by a graph $G$ which consists of nodes $N$ and edges $E$ in Equation (1). The graph $G$ is shown in Figure 1. The nodes represent intersections and the edges represent road segments between intersections. In this regard, all road segments are represented as two oppositely-oriented edges. $e_{ij}$ is an edge whose starting point is $n_i$ and end point is $n_j$.

$$G = (N, E), N = \{n_1, n_2, ...\}, E = \{e_{ij} | n_i, n_j \in N\} \tag{1}$$

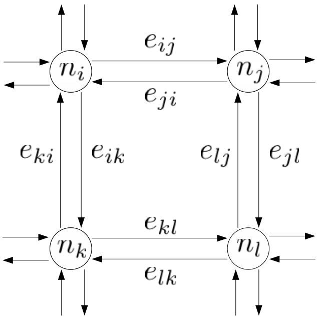


**Fig. 1.** A road network in a travel area

The edge $e_{ij}$ is given by Equation (2). The probability $\alpha_{ij}$ is the expected-value of customers' existence on $e_{ij}$ and $U_{ij}$ is the customer list. A customer appears on $e_{ij}$ with $\alpha_{ij}$ per unit time and is added to $U_{ij}$.

$$e_{ij} = (\alpha_{ij}, U_{ij}) \tag{2}$$

For simplicity, the time required for vehicles to go through edges is defined to be constant and also the time required to turn right or left at intersections is not considered.

We define an edge whose $\alpha_{ij}$ exceeds the threshold $\epsilon$, as a high expectation edge. High expectation edges are given by $E_h$ in Equation (3).

$$E_h = \left\{ e^h_{ij} = e_{ij} | \alpha_{ij} > \epsilon \right\} \subseteq E \tag{3}$$

The customer $u_m$ is given in Equation (4) which contains an appearance edge $e^a_m$, a destination node $n^d_m$, and a waiting time $t^w_m$. A customer $u_m$ waits on $e^a_m$ until the waiting time $t^w_m$ elapse since appearance of $u_m$. If any vehicles get to $e^a_m$ before $t^w_m$ elapsed, a customer $u_m$ conveys a destination node $n^d_m$ and order the driver so as to carry $u_m$ to $n^d_m$. Then, $u_m$ is removed from the customer list $U_{ij}$.

$$u_m = \left( e^a_m, n^d_m, t^w_m \right) \tag{4}$$

Vehicles are given by $V$ in Equation (5) and a vehicle $v_i$ is given in Equation (6) which contains a cluster $c_i$, route $r_i$, and destination node $n^d_i$. A cluster $c_i$ is given in Equation (7) and is a set of high expectation edges which are assigned to the vehicle $v_i$. The vehicle $v_i$ drives on the edges in the route $r_i$, which is given in Equation (8). When the vehicle $v_i$ passes the edge $e_{ij}$, if any customers exist in the customer list $U_{ij}$, $v_i$ gains the customer $u_m$ from the head of $U_{ij}$. Then, $n^d_i$ is substituted with $n^d_m$, which is the destination node of $u_m$ and the shortest path to $n^d_i$ is assigned to the route $r_i$.

$$V = \{v_0, v_1, ..., v_K\} \tag{5}$$

$$v_i = \left( c_i, r_i, n^d_i \right) \tag{6}$$

$$c_i = \{e^c_{i0}, e^c_{i1}, ..., e^c_{il}\} \tag{7}$$

$$r_i = \{e^r_{i0}, e^r_{i1}, ..., e^r_{im}\} \tag{8}$$

## 4 Travel Route Assignment

In this section, we propose the method for travel route assignment. Henceforth, we denote the distance between edges $e_{ij}$ and $e_{kl}$ as $d(e_{ij}, e_{kl})$. Since the time required for vehicles to go through edges is constant, $d(e_{ij}, e_{kl})$ is the number of edges in the shortest path from node $n_j$ to node $n_k$.

### 4.1 Setting Primary Routes

Initially, K edges are assigned to K vehicles as their start edge $e_i^s$ in descending order of $\alpha_i$. Then, $e_i^s$ of each vehicle $v_i$ is added to cluster $c_i$ and route $r_i$. In addition, if $e_i^s$ is an element of high expectation edges $E_h$, $e_i^s$ is removed from $E_h$. Then, primary clusters and routes are constructed and assigned to each vehicle as follows.

```
Construction of primary clusters and routes
while $E_h != null$ do
    pick up $(v_i, e_{jk}^h)$ whose $d(e_i^a, e_{jk}^h)$ is smaller than any other pair $(e_x^a, e_{yz}^h)$;
    add $e_{jk}^h$ to $C_i$;
    add the local path from $e_i^a$ to $e_{jk}^h$ into $r_i$;
    $e_i^a \leftarrow e_{jk}^h$;
    remove $e_{jk}$ from $E_h$;
end while;
for each $v_i \in V$ do
    add the local path from $e_i^a$ to $e_i^s$ into $r_i$;
end for
```

A local path between two high expectation edges is input by the method based on Dijkstra algorithm[5]. The typical Dijkstra algorithm is used about the shortest route finding problem, while we utilize it as a means to find a local path which has the largest average expected-value of customers' existence in the shortest path from the start edge $e_s$ to the end edge $e_e$. In our method, the average expected-value of customers' existence is substituted for the length of the edge. The attention node $n_a$ is selected in descending order of the distance from the end point node of start edge. Then, the maximum average expected-value of customers' existence in the shortest path from the end point of $e_s$ to the attention node $n_a$ is appended to $n_a$ as $E_{max}$. When $E_{max}$ is appended to the start point node of $e_e$, we achieve the optimal local path. An example is shown in Figure 2. Circles represent nodes, arrows represent edges, numbers appended to arrows represent expected-values of customers' existence, numbers in circles represent $E_{max}$ of the nodes, and thick arrows represent high expectation edges. $n$ is the end point node of $e_s$ and $n'$ is the start node of $e_e$.

### 4.2 Reassembly of Clusters

In order to maintain the optimal assignment of pathways, vehicles exchange pathways mutually in their clusters. We define the membership value to convert

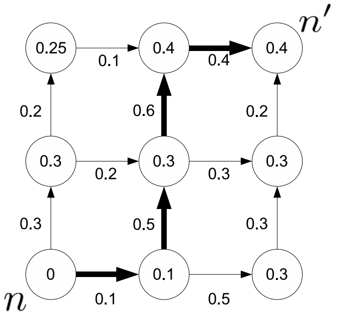


**Fig. 2.** Example of setting a local path

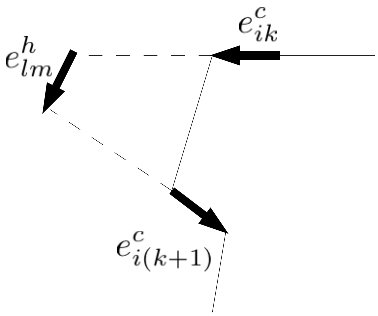


**Fig. 3.** Example of route change

clusters of vehicles. The membership value of edge $e_{jk}$ to the cluster $c_i$ is larger if the change of length of the route $r_i$ by converting $c_i$ is smaller. Consider the change of route $r_i$ to pass the high expectation edge $e^h_{lm}$ between edge $e^c_{ik}$ and $e^c_{i(k+1)}$ as shown in Figure 3, for example. In this instance, the preliminary membership value of edge $e^h_{lm}$ to the cluster $c_i$: $m'(e_{lm}, c_i, k)$ is expressed in Equation (9).

$$m'(e^h_{lm}, c_i, k) = \frac{1}{d(e^c_{ik}, e^h_{lm}) + d(e^h_{lm}, e^c_{i(k+1)}) - d(e^c_{ik}, e^c_{i(k+1)})} \tag{9}$$

In all $m'(e^h_{lm}, c_i, k)$ with k which is neither less than 0 nor more than m, the largest one is assigned to the membership value $m(e^h_{lm}, c_i)$.

When a vehicle $v_i$ gains a customer, all of high expectation edges in the cluster $c_i$ are removed from $c_i$ and added to $E_h$. At the predetermined update interval, clusters are reassembled as follows. First, each high expectation edge $e_{jk}$ in $E_h$ is added to the cluster $c_i$ which the membership value $m(e_{jk}, c_i)$ is larger than membership values of $e^h_{jk}$ to any other clusters. Second, as for each vehicle $v_i$ whose cluster is empty, the high expectation edge $e^h_{lm}$, which is nearest from destination node $n^d_i$, changes its belonging cluster to $c_i$. Then, assuming that each high expectation edge $e^h_{jk}$ changes its belonging cluster to $c_i$ in the descending order of membership value $m(e^h_{jk}, c_i)$, clusters are converted actually if the condition expressed in Equation (10) is satisfied. In Equation (10), $|r_i|$ represents the length of route $r_i$, $P_{bef}$ and $P_{aft}$ represent local paths before and after the cluster change, respectively, and $|P_{bef}|$ and $|P_{aft}|$ represent the lengths of them.

$$\frac{\Sigma_{e_k \in P_{bef}} \alpha_k}{|r_i|} > \frac{\Sigma_{e_k \in P_{aft}} \alpha_k}{|r_i| - |P_{bef}| + |P_{aft}|} \tag{10}$$

By converting clusters with the above method, clusters and routes of some other vehicles are changed. Therefore, in order to maintain the optimal travel routes as a whole, the above method is applied to all vehicles whose cluster is affected by above method. The method continues to be applied recursively until clusters stop changing.

# 5 Experiment

## 5.1 Models of Driving Taxis

In order to assess the validity of our method, we compare the taxi model based on our method with the models of existing driving taxis. The models of driving taxis decide their travel routes individually. In this paper, two models are defined: Successive Routing (SR) model and Nearest High Expectation edge Select (NHES) model.

**Successive Routing Model.** Vehicles of SR model select an edge with largest expected-value of customers' existence in edges they can pass next at each intersection.

**Nearest High Expectation Edge Select Model.** Vehicles of NHES model search the nearest high expectation edge and set it as destination edge $e_d$. Vehicles which are not carrying a customer are heading in the direction of $e_d$ constantly.

## 5.2 Parameter Setting

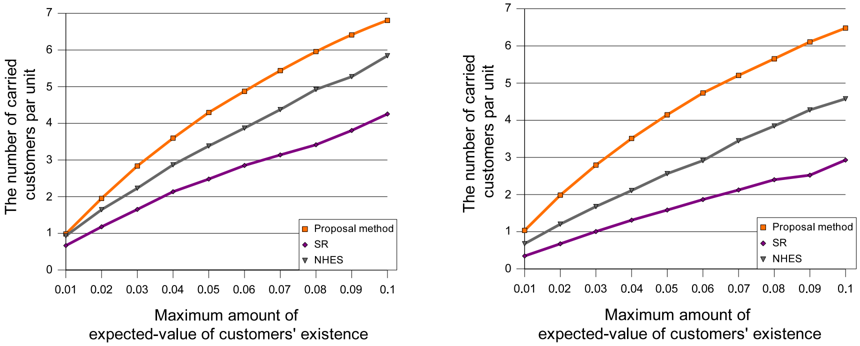
A road network in the traveling area is set to a grid network ($20 \times 20$). The unit time is represented by $T$ and the test duration of one trial is $200 \times T$. The edge length is represented by $|e|$ and the traveling speed of vehicles is $|e|/T$ consistently. For simplicity, the distance between the appearance edge $e_m^a$ and the destination node $n_m^d$ of all customers is defined as $10 \times |e|$. The update interval of clusters is $5 \times T$ and the wait time $t_m^w$ of all customers is $5 \times T$. The maximum amount of expected-value of customers' existence is represented by $\alpha_{max}$ and the threshold $\epsilon$ for selecting high expectation edges is set to $\frac{7}{10} \times \alpha_{max}$.

Expected-values of customers' existence $\alpha_{ij}$ are appended to each edge $e_{ij}$ based on random numbers. In order to survey effect of customers' appearance trend, we ran experiments with two patterns of appending $\alpha_{ij}$. In pattern 1, uniform random numbers ($0 \sim \alpha_{max}$) are appended as $\alpha_{ij}$. In pattern 2, squares of uniform random numbers ($0 \sim \alpha_{max}$) are appended as $\alpha_{ij}$.

## 5.3 Result of Experiments

In our experiments, vehicles which utilize our method, SR model and NHES model traveled at the same time and area. All results of experiments are the averages of 100 trials.

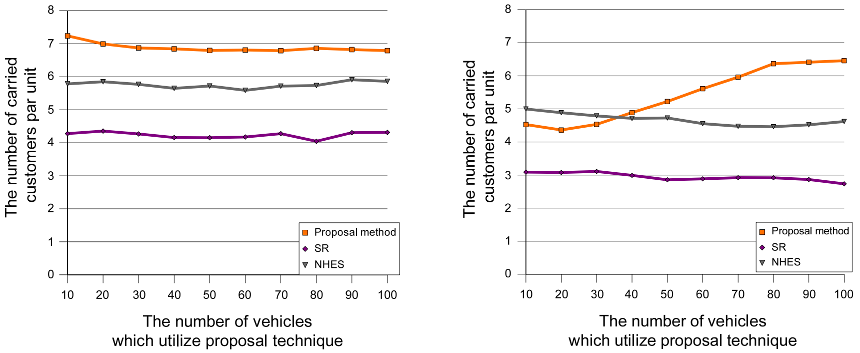
First, we report experimental results related to expected-value of customers' existence. The maximum amount of expected-value of customers' existence $\alpha_{max}$ is set from 0.01 to 0.1. The number of vehicles in each model is set to a fixed value 100. As Figure 4 shows, vehicles which utilize our proposal method gain more customers than any other models constantly. In addition, the results of 2 other models in pattern 2 are considerably fewer than that of pattern 1, but the result of vehicles which utilize the proposal method in pattern 2 is close to that in pattern 1.



(a) Pattern 1

(b) Pattern 2

**Fig. 4.** Experimental results related to expected-value of customers' existence



(a) Pattern 1

(b) Pattern 2

**Fig. 5.** Experimental results related to the number of vehicles

Second, we report experimental results related to the number of vehicles. The number of vehicles which utilize our proposal method is set from 10 to 100 and the number of vehicles of 2 other models is set to a fixed value 100. The maximum amount of expected-value of customers' existence is set to 0.1. As Figure 5 shows, in pattern 1 results are nearly-constant regardless of the number of vehicles and the result of proposal method is better than results of other models. In pattern 2, if the number of vehicles is fewer than 40, the result of proposal method is worse than the result of NHES model. If the number of vehicles is greater than 80, the result is close to that in pattern 2.

As a result, it is thought that our proposal method can reduce the effects of customers' appearance trend if the number of vehicles takes over a certain number.

## 6 Conclusion

In this paper, we focused on the routing problem of cruising taxis and proposed the method to maintain their optimal travel routes. In order to gain customers effectively, cruising taxis must travel concertedly. Our method makes an adaptive routing of them and their coordinated traveling possible. The simulation results show that our method is more efficient than the existing cruising taxis which travel individually. However, in order to apply our method to a real world, we must assume the existence of some other cruising taxis which utilize other coordinated methods. In our future work, we must consider other algorithm for routing the cruising taxis and compare it with our method in this paper.

## References

1. Liao, Z.: Taxi dispatching via global positioning systems. IEEE Transactions on Engineering Management 48(3), 342–347 (2001)
2. Bektas, T.: The multiple traveling salesman problem: an overview of formulations and solution procedures. Omega: The International Journal of Management Science 34(3), 209–219 (2006)
3. Stutz, C., Runkler, T.A.: Classification and prediction of road traffic using application-specific fuzzy clustering. IEEE Transactions on Fuzzy Systems 10(3), 297–308 (2002)
4. Bezdek, J.C., Ehrlich, R., Full, W.: Fcm: The fuzzy c-means clustering algorithm. Computers & Geosciences 10(2-3), 191–203 (1984)
5. Johnson, D.B.: A note on dijkstra's shortest path algorithm. J. ACM 20(3), 385–388 (1973)

# Route Optimization Using Q-Learning for On-Demand Bus Systems

Naoto Mukai[1], Toyohide Watanabe[2], and Jun Feng[3]

[1] Department of Electrical Engineering,
Tokyo University of Science
Kudankita, Chiyoda-ku, Tokyo, 102-0073, Japan
mukai@ee.kagu.tus.ac.jp
[2] Department of Systems and Social Informatics,
Graduate School of Information Science, Nagoya University
Furo-cho, Chikusa-ku, Nagoya, 464-8603, Japan
Department of Electrical Engineering
watanabe@is.nagoya-u.ac.jp
[3] Hohai University, Nanjing, Jiangsu 210098, China
fengjun-cn@vip.sina.com

**Abstract.** In this paper, we focus on a new transport service called on-demand bus system. A major feature of the system is that buses pick up customers door-to-door when needed or required. Thus, there is no pre-determined travel routes for buses, and travel routes must be changed according to the occurrence frequency of customers. In order to find a more effective travel plan to the problem, we adopt Q-learning which is one of the machine learning algorithms. However, native Q-learning is inadequate to our target problem because the number of customers at pick-up points is time-dependent. Therefore, we improve an update process of Q values and a selection process of the next pick-up point, on the basis of time passage parameters. In particular, rewards are understated in update process, on the other hand, Q values are overstated in selection process. At the last, we report our simulation results and show the effectiveness of our algorithm for the problem.

## 1 Introduction

In these years, some local communities adopt a new transport service called on-demand bus system in stead of a fixed bus system in Japan [1,2,3]. The on-demand bus system is more cost efficient than traditional transport services because buses pick up customers door-to-door when needed or required. Thus, there is no pre-determined travel routes for buses, and travel routes must be changed according to the occurrence frequency of customers. This problem can be regarded as one of the VRPs (Vehicle Routing Problems) and its variants [4,5]. The VRP and its variants are combinational optimization problems, and meta-heuristic algorithms such as GA (Genetic Algorithm) are major approaches to the problems [6,7]. However, such meta-heuristic algorithms are unfit to dynamic problems such as the on-demand bus problem. Therefore, in this paper, we

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 567–574, 2008.

adopt Q-learning algorithm [8,9] which is one of the bootstrap machine learning algorithms to find an effective travel plan. Q-learning updates a value of a rule (i.e., sum of discounted rewards) by a behavioral experience in Markov decision process, and can find the optimal solution subject to appropriate learning and discount rates in an infinite time. However, native Q-learning is inadequate to our target problem because rewards (i.e., customers) depend on the passage of time. This issue make it difficult to converge of a solution in a learning process. Therefore, we improve an update process of Q value and a selection process of the next node, on the basis of time passage parameters called elapsed time, elapsed distance, and elapsed cost. In particular, rewards are understated in update process, on the other hand, Q values are overstated in selection process.

The remainder of this paper is as follows: Section 2 defines a problem setting for the on-demand bus problem. Section 3 proposes a route optimization using Q-Learning for our target problem. Section 4 reports our experimental results. Finally, Section 5 offers our conclusions and future works.

## 2 Problem Setting

In this section, we formalize the on-demand bus problem. There are two types of on-demand bus systems: semi-demand and full-demand. In semi-demand type, a request of a customer is a travel from a designated place (such as a station) to any point (such as a home) or a travel from any point to a designated place. On the other hand, in full-demand type, a request of a customer is a travel from any point to any point. In this paper, we focus on semi-demand type.

A problem space for semi-demand type is represented by a directed graph as shown in Figure 1 [1]. A node $n$ represents a pick-up point, and an edge $e$ represents a link between two nodes [2]. In the graph, there are two special nodes: start node $S$ and goal node $G$ of a travel route. In particular, a vehicle starts from $S$, and picks up customers along a travel route through free node choice, and drops off customers at $G$.

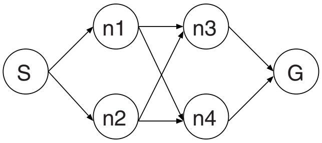


**Fig. 1.** A sample of a graph

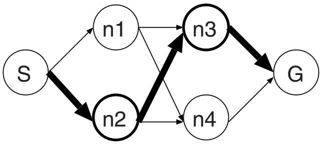


**Fig. 2.** Travel route

A update formula of the number of customers for each node is defined as Equation 1 where $r(n)_t$ is the number of customers at time $t$ for node $n$, and $\mu(n)$ is an increasing number for node $n$. The increasing number $\mu(n)$ represents an occurrence frequency of customers for node $n$, but this value is unknown

[1] Allowing travel in only one direction.

[2] Traveling time between nodes is ignored.

**Table 1.** Elapsed time

| **Travel** | $\xi(n_1)$ | $\xi(n_2)$ | $\xi(n_3)$ | $\xi(n_4)$ |
|---|---|---|---|---|
| before | 0 | 1 | 0 | 1 |
| after | 1 | 0 | 0 | 2 |

for drivers. Hence, drivers must estimate more accurate value of $\mu(n)$ based on their transport histories to increase their rewards (i.e., the number of pick-up customers). Moreover, based on the estimated value of $\mu(n)$, vehicles can select profitable travel routes according to time $t$.

$$r(n)_{t+1} = r(n)_t + \mu(n) \tag{1}$$

If a vehicle does not visit a node (i.e., a vehicle selects another node), the customers at the node must wait for next arriving of the vehicle. Moreover, due to a capacity $\nu$ of a vehicle (i.e., the maximum amount that can be picked), the maximum reward for a vehicle is equal to $\nu$ at one travel from $S$ to $G$ as shown in Equation 2 where $\sum r(n)$ is the sum of reward in one travel. Therefore, a driver must change its travel routes in a cycle according to the passage of time. As mentioned before, native Q-learning cannot adjust such cyclic process. Thus, we improve update and selection processes of Q-learning in the next section.

$$\sum r(n) \leq \nu \tag{2}$$

Here, we define elapsed time $\xi(n)$ for node $n$. A elapsed time $\xi(n)$ means a passage of time from when a vehicle visits node $n$. For example, if a travel route of a vehicle is as shown in Figure 2, the elapsed times for the graph is changed as Table 1. Note that $\xi(n_1)$ and $\xi(n_4)$ are incremented because the vehicle did not visit the nodes. On the other hand, $\xi(n_2)$ and $\xi(n_3)$ are reset because the vehicle visited the nodes.

Moreover, we define elapsed distance $\tau(n)$ and elapsed cost $c(n)$ for node $n$. A elapsed distance $\tau(n)$ means the sum of the elapsed time in a path from $n$ to $G$, and an elapsed cost $c(n)$ is the number of visit nodes in the path. For example, if a travel route of a vehicle is as shown in Figure 2, the elapsed distances for the graph is changed as Table 2. Here, we focus on $n_1$ after the travel. There are two paths from $n_1$ to $G$: $n_1 \rightarrow n_3$ or $n_1 \rightarrow n_4$ as shown in Equations 3 and 4. Consequently, elapsed distance $\tau(n_1)$ is the maximum value among the paths $(= 3)$ as shown in Equation 5, and elapsed cost $c(n_1)$ is the number of visit nodes $(= 2)$.

**Table 2.** Elapsed distance

| **Travel** | $\tau(n_1)$ | $\tau(n_2)$ | $\tau(n_3)$ | $\tau(n_4)$ |
|---|---|---|---|---|
| before | 1 | 2 | 0 | 1 |
| after | 3 | 2 | 0 | 2 |

$$\tau(n_1 \rightarrow n_3) = \xi(n_1) + \xi(n_3) = 1 \tag{3}$$

$$\tau(n_1 \rightarrow n_4) = \xi(n_1) + \xi(n_4) = 3 \tag{4}$$

$$\tau(n_1) = \max(\tau(n_1 \rightarrow n_3), \tau(n_1 \rightarrow n_4)) = 3 \tag{5}$$

## 3 Route Optimization Using Q-Learning

A state and an action of Markov decision process corresponds to a node and an edge of the graph in our target problem. Moreover, a reward $r(n)$ for a driver corresponds to the number of pick-up customers at node $n$. The Q value represents the value of action $e$ at state $n$, i.e., the sum of expected discounted rewards. A driver selects the next visit node in the graph by using softmax method. The softmax method provides selection probabilities of nodes at branch points on the basis of a ratio of Q values. We explain an update process of Q values and a selection process of visit nodes as follows.

### 3.1 Update Process

The update function of native Q-learning is defined as Equation 6, where $Q(n,e)$ represents a value when a driver selects edge $e$ at node $n$, $Q(n',e')$ represents a value when a driver selects edge $e'$ at the next visit node $n'$ from $n$, $E(n')$ represents a set of edges at node $n'$, $\alpha$ is a learning rate, and $\gamma$ is a discount rate. The learning rate $\alpha$ adjusts the priority balance between present rewards and estimated Q values. The discount rate shows the priority of rewards in the future.

$$Q(n,e) \leftarrow Q(n,e) + \alpha \left[ r(n) + \gamma \max_{e' \in E(n')} Q(n',e') - Q(n,e) \right] \tag{6}$$

The update function of improved Q-learning for our target problem is defined as Equation 7. Note that $r(n)$ is replaced by $r(n)/(\xi(n)+1)$. It means that reward $r(n)$ is divided by elapsed time $\xi(n)$. Consequently, the calculated value represents reward for a unit of time because reward $r(n)$ increases with time.

$$Q(n,e) \leftarrow Q(n,e) + \alpha \left[ \frac{r(n)}{\xi(n)+1} + \gamma \max_{e' \in E(n')} Q(n',e') - Q(n,e) \right] \tag{7}$$

### 3.2 Selection Process

The selection probability of native softmax method is defined as Equation 8, where $Q(n,e)$ represents a value when a driver selects edge $e$ at node $n$, $E(n)$ represents a set of edges at node $n$, and $T$ represents a temperature parameter of Boltzmann distribution. The temperature parameter $T$ is decreased with time passage as shown in Equation 9, where $\psi$ is a weight parameter. Hence, when the temperature $T$ is high, the native softmax method randomly selects the next node independently of Q value. On the other hand, when the temperature $T$ is sufficiently-small, the native softmax method selects a high Q-value node as the next node with a high probability.

$$p(e|n) = \frac{\exp^{Q(n,e)/T}}{\sum_{e \in E(n)} \exp^{Q(n,e)/T}} \tag{8}$$

$$T_{t+1} = \psi \times T_t \qquad (0 < \psi < 1) \tag{9}$$

The selection probability of improved softmax method is defined as Equation 10. Note that $Q(n, e)$ is weighted by $w(n)$ as shown in Equation 11. The weight function $w(n)$ represents extra reward according to the ratio of elapsed distance $\tau(x)$ and elapsed cost $c(x)$. When the temperature $T$ is sufficiently-small, the native softmax method selects the same high Q-value node without the significant changes of Q values. On the other hand, the improved softmax method can change its selection in a cycle according to $\tau(x)$ and $c(x)$ even when the temperature $T$ is sufficiently-small.

$$p(e|n) = \frac{\exp^{w(n) \times Q(n,e)/T}}{\sum_{e \in E(n)} \exp^{w(n) \times Q(n,e)/T}} \tag{10}$$

$$w(n) = 1 + \frac{\tau(x)}{c(x)} \tag{11}$$

# 4 Experiment

In this section, we report our experimental results to evaluate our improved Q-learning algorithm for on-demand bus problem by using a computer simulation.

## 4.1 Experimental Setting

In our simulation, we used two graphs as shown in Figures 3 and 4. In both the graphs, there are 11 nodes. A start node is $d_0$ and a goal node is $n_9$. The occurrence rates of customers at nodes are set from 0 to 9, randomly. A difference between the two graphs is a topology: there are $3 \times 3 = 9$ paths between start

**Table 3.** Parameter setting

| Parameter | Value |
|---|---|
| Maximum Cycle | 1000 |
| Maximum Capacity $\nu$ | 50 |
| Occurrence Rate $\mu$ | $0 - 9$ |
| Initial Temperature $T$ | 100 |
| Temperature Weight $\psi$ | 0.99 |
| Learning Rate $\alpha$ | 0.01 |
| Discount Rate $\gamma$ | 0.8 |

and goal nodes in graph 1, and only 3 paths in graph 2. In one simulation cycle, a vehicle starts from $d_0$, and picks up customers along a travel route according to selection probabilities, and drops off customers at $n_9$. This simulation cycle is repeated 1000 times. We compare three transport strategies: MIX[3], Q-Learning with softmax, and improved Q-learning with softmax. A parameter setting is summarized in Table 3.

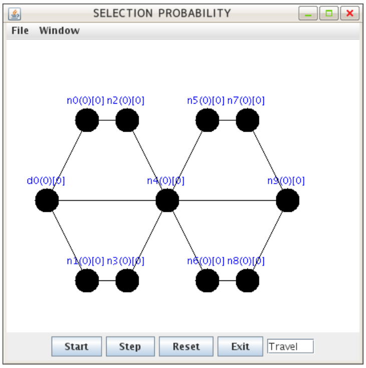


**Fig. 3.** Graph 1 (9 paths)

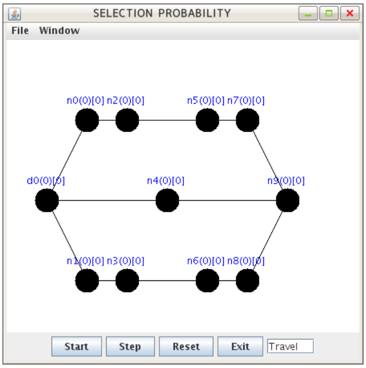


**Fig. 4.** Graph 2 (3 paths)

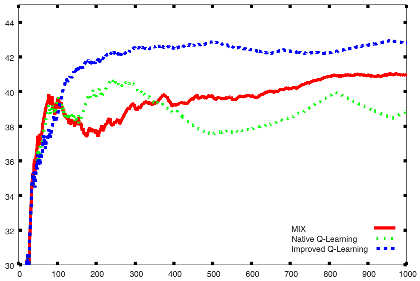


**Fig. 5.** Average (Graph 1)

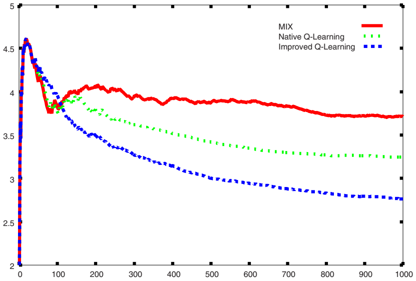


**Fig. 6.** Standard Deviation (Graph 1)

## 4.2 Experimental Results

An averages of reward (the number of pick-up customers) for graph 1 and graph 2 are illustrated in Figures 5 and 7. A standard deviation of reward (the number of pick-up customers) for graph 1 and graph 2 are illustrated in Figures 6 and 8. From the results, we found that the improved Q-learning algorithm can keep high reward. On the other hand, the native Q-learning algorithm never converge at the last of the simulation cycle. This fact can be explained as follows. The improved algorithm change the selection probabilities of the next node in a cycle even when the temperature $T$ is sufficiently-small. On the other hand, the native algorithm selects a high Q-value node

[3] A driver selects the next node from neighbors in equal probability.

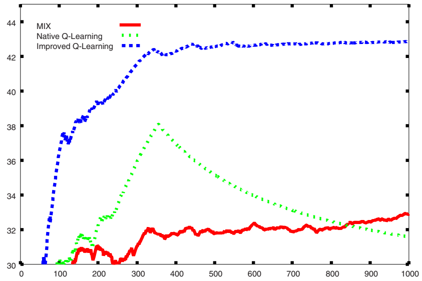


**Fig. 7.** Average (Graph 2)

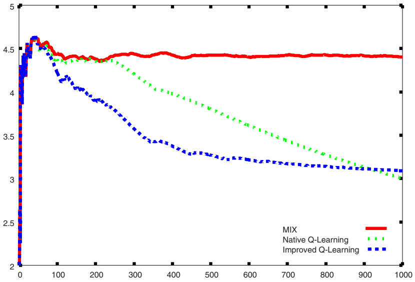


**Fig. 8.** Standard Deviation (Graph 2)

with high probability. Thus, the customers at a low Q-value node are left for a long time until the next arriving of vehicles, and the reward of the low Q-value node rises temporarily in the next cycle. Consequently, these results indicate that our improved Q-learning algorithm helps to increase the reward for drivers in on-demand bus systems compared to the native Q-learning and the Mix strategies.

## 5 Conclusion

In this paper, we focused on the on-demand bus problem. The on-demand bus system is a new type of transport service, and the travel routes for the buses are not determined in advance. In order to solve the problem, we proposed a new algorithm based on Q-learning to increase the profit of drivers. In the algorithm, rewards are understated on the basis of the elapsed time in update process, and Q values are overstated based on the elapsed distance and cost in selection process. Our simulation result indicate that the improved Q-learning algorithm outperforms other strategies (the native Q-learning and Mix strategy). As for a future work, we have to consider a competition or a cooperative behaviors among drivers on the basis of Q values.

## Acknowledgment

This work was supported by Grant-in-Aid for Young Scientists Start-up (19810022).

## References

1. Ohta, M., Shinoda, K., Noda, I., Kurumatani, K., Nakashima, H.: Usability of demand-bus in town area. Technical Report 2002-ITS-11-33. Technical Report of IPSJ (2002) (in Japanese)
2. Noda, I., Ohta, M., Shinoda, K., Kumada, Y., Nakashima, H.: Is demand bus reasonable in large scale towns? Technical Report 2003-ICS-131. Technical Report of IPSJ (2003) (in Japanese)

3. Harano, T., Ishikawa, T.: On the validity of cooperated demand bus. Technical Report 2004-ITS-19-18. Technical Report of IPSJ (2004) (in Japanese)
4. Desrochers, M., Lenstra, J., Savelsbergh, M., Soumis, F.: Vehicle routing with time windows: Optimization and approximation. Vehicle Routing: Methods and Studies, pp. 65–84 (1988)
5. Solomon, M., Desrosiers, J.: Time window constrained routing and scheduling problems. Transportations Science 22, 1–13 (1988)
6. Kanoh, H., Nakamura, N., Nakamura, T.: Route selection with unspecified sites using knowledge based genetic algorithm. The Transactions of the Japanese Society for Artificial Intelligence 17, 145–152 (2002)
7. Keiichi, U., Ryuji, M.: A real-time dial-a-ride system using dynamic traffic information. The Transactions of the Institute of Electronics, Information and Communication Engineers. A 88, 277–285 (2005)
8. Watkins, C.J.C.H., Dayan, P.: Technical note: Q-learning. Machine Learning 8, 55–68 (1992)
9. Jaakkola, T., Jordan, M.I., Singh, S.P.: On the convergence of stochastic interactive dynamic programming algorithms. Neural Computation 6, 1185–1201 (1994)

# Sharing Event Information in Distributed Environment Based on Three-Layered Structure

Masakazu Ikezaki[1], Toyohide Watanabe[1], Tomoko Kojiri[1],
and Taketoshi Ushiama[2]

[1] Department of Systems and Social Informatics,
Graduate School of Information Science, Nagoya University,
Furo-cho, Chikusa-ku, Nagoya, 464-8603, Japan
{mikezaki,watanabe,kojiri}@watanabe.ss.is.nagoya-u.ac.jp
[2] Faculty of Design, Kyushu University,
9-1 Shiobaru 4-chome, Minami-ku, Fukuoka, 815-0032, Japan
ushiama@design.kyushu-u.ac.jp

**Abstract.** Recently, geographic information models for handling events have been proposed. However, the models are not always suitable to handle broadly spreading events, specifically in a distributed environment. This is because event data should be synchronized between geographic objects. In this paper, we propose a three-layered structure model with a view to handling events successfully in the distributed environment. In our model, event data and geographic objects are represented independently in the 1st layer, and they are associated dynamically according to a query in the 2nd layer. In the 3rd layer, data propagated from the 2nd layer are aggregated so as to be used in application systems. Based on this framework, event data and geographic objects could be represented independently, and event data computed in the 2nd layer would be shared and re-used.

**Keywords:** GIS, Event, Spatio-temporal Information.

## 1 Introduction

Recently, in Geographic Information Systems (GISs), models for handling events[1],[2]. and frameworks for sharing geographic information over the network[3],[4] are hottest topics, respectively. However, it is difficult to apply the traditional models to an event handling system in distributed environment. This is because an event is inherently aggregation of a number of related geographic changes; hence, it is represented as a tuple of geographic objects changed by the event. In other words, management of geographic objects and events data costs so much, because they need to be synchronized each other.

We have worked on the framework for treating event information[5][6]. In [7], we proposed a model for associating events and related geographic objects in the distributed environment. In our model, geographic objects and events are managed independently, and associations among them are computed according to a query for an event. In this paper, we propose a three-layered structure model based on our data association model, in order to realize share of event data. By using this framework, event data could be used from other application systems in the distributed environment. Here, the application systems mean GISs that use event information.

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 575–582, 2008.

This paper is organized as follows. In Section 2, we show a framework of three-layered model. In Section 3, some computing examples based on our proposed model are provided. The discussion for our model is set in Section 4. Finally, in Section 5 we state our conclusion and future works.

## 2 Three-Layered Structure

The process for preparing data used in application systems is as follows. (1) Gathering or making geographic objects, (2) analyzing and computing event data from the geographic objects or gathering it from another data source, (3)associating event with geographic objects related implicitly to events based on the analysis of geographic objects, and (4) representing geographic objects and event features based on this association. This process is needed to develop application systems, whether the systems are based on a stand-alone architecture or a distrubuted architecture. Here, event data and geographic objects in the processes 1, 2 are closed in types of events and domains of geographic objects. On the other hand, in the process 3, to associate events with geographic objects needs to handle geographic objects over the domains. In addition, the data representation in the process 4 depends on an application system. Therefore, in order to realize data sharing, we develop a multi-layer structure corresponding to the processes. Namely, by dividing the data into several layers, developers of application systems could re-use the data that were gathered or calculated by another developer.

Fig.1 shows our framework . A proposed model consists of three layers. The 1st layer is the data management layer in which event data, geographic objects and meta information for sharing information are stored distributedly, from one domain to another. The 2nd layer is the data association layer. Here, procedures for associating an event with related geographic objects are stored. By using the procedures, geographic object sets related to an event are computed. To store only the procedures makes it possible to update event data and geographic objects without synchronization in each domain. In the 3rd layer, procedures for deriving event features used in application systems are stored. This is the event representation layer. Through the 2nd and 3rd layers, data set managed in distributed databases can be transformed into a data representing event features used in application systems. By using the stored procedures in the 2nd layer and 3rd layer, event data sharing and reutilization among a number of application systems would be realized.

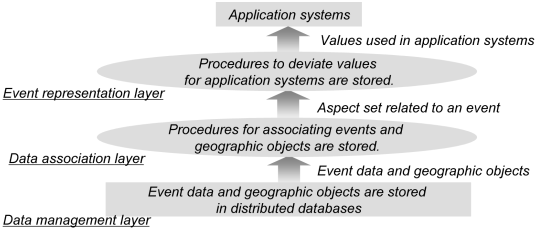

Application systems
Values used in application systems
Procedures to deviate values
for application systems are stored.
Event representation layer
Aspect set related to an event
Procedures for associating events and
geographic objects are stored.
Data association layer
Event data and geographic objects
Event data and geographic objects are stored
in distributed databases
Data management layer

**Fig. 1.** Three-layered structure

### 2.1 1st Layer: Data Management Layer

In the 1st layer, geographic objects and event data are managed distributedly. The management units are supposed to be local governments, government agencies and so on. In this paper, simplistically, the management units would correspond to types of events and domains of geographic objects. In other words, all of geographic objects and events could be reached by specifying the types. In addition, central control mechanisms such as access transparency or location transparency are needed in order to reach each data. However, they would also be omitted for simplification.

In spatio-temporal GISs, a geographic object is an entity with the shape and lifespan, and attribute values of a geographic object are changeable during the lifespan. An event is represented by attributes and occurrence fields in the space and time. The occurrence fields represent the histories about the existence space of an event. Geographic objects $o$ and events $ev$ are represented as follows.

$$\begin{aligned} o &= (t_s, t_e, ATTR_o, POS_o), \\ ATTR_o &= \{\{(val_{i_0}, t_s, t_{i_1}), ..., (val_{i_{m_i-1}}, t_{i,m_i-1}, t_e)\} | i = 0, \ldots, n\}, \\ POS_o &= \{(pos_0, t_s, t_1), ..., (pos_{l-1}, t_{l-1}, t_e)\}. \end{aligned} \tag{1}$$

$$\begin{aligned} ev &= (t_s, t_e, ATTR_e, POS_e), \\ ATTR_e &= (val_0, ..., val_{n-1}), POS_e = \{(pos_0, t_s, t_1) \ldots (pos_{m-1}, t_{m-1}, t_e)\}. \end{aligned} \tag{2}$$

Here, $t_s$ and $t_e$ are the onset time and termination time, respectively. $ATTR_o$ is a set of attribute values of $o$. Attributes of $o$ are represented as a set of tuples: the attribute values and the valid time intervals. A number $n$ is a number of the attributes and a number $m_i$ is the number of the attribute value's histories. $POS_o$ represents the existence region of $o$, and each $pos_i$ represents the existence space with the valid time. $ATTR_e$ is a tuple of invariant event's attributes, corresponding to event types. $POS_e$ represents the occurrence field of $ev$, and each $pos_i$ is a spatial region of the event with the valid time.

### 2.2 2nd Layer: Data Association Layer

In the 2nd layer, sets of geographic objects related to events are transfered to the 3rd layer as aspect sets. An aspect represents a focused time period in a life span of a geographic objects. The data stored in the 2nd layer are not aspect sets, but procedures to compute the aspect sets. Therefore, the associations could be computed independently from changes of data in the 1st layer. The procedures are based on two operations:the specification of geographic objects that have interesting changes or that participate in events, and set operation among geographic object sets.

**Aspect of geographic objects.** We call an ordered set of attribute values in an object as geographic object's behavior. By indicating a pattern of an interesting geographic object's behavior, a set of geographic objects which have the interesting behavior can be specified. The behavioral pattern *bpt* is described as follows.

$$bpt = \{(classname_i, state+) | i = 0, \ldots n\}, state = (attname, cond). \tag{3}$$

Here, $classname$ is a name of geographic object class, and $state$ represents the state of a geographic object. The signature $+$ means more than one iteration. $state$ is represented as a tuple of an attribute name $attname$ and a condition of the attribute value $cond$. As an example of collapse of buildings and roads, if the class names and attributes of buildings and roads are $Buildings$ and $strength$, $Roads$ and $state$, respectively, then the behavioral pattern is described as

$$\begin{aligned} bpt_{collapse} &= \{(Buildings, < st_1, st_2 >), (Roads, < st_3, st_4 >)\}, \\ st_1 &= (strength, 6 \leq val < 10), st_2 = (strength, 1 \leq val < 5), \\ st_3 &= (state, val = normal), st_4 = (state, val = destructed). \end{aligned} \tag{4}$$

Here, the notation $<>$ represents an ordered list on the time axis and $val$ represents the attribute value corresponding to $attname$. In this example, the "collapsed" behavior of $Buildings$ and $Roads$ is represented as the changes of buildings from a state, whose $strength$ value is 6 to 10, to another state, whose $strength$ value is 1 to 5, and the changes of roads from a $normal$ state to a $destructed$.

The aspect of a geographic object is represented by a tuple of a geographic object and a time interval corresponding to the behavior.

$$aspect = (obj, T), start(obj) < t \in T < end(obj). \tag{5}$$

Here, $obj$ represents the geographic object. $T$ is a set of instant times, and represents the focused time interval. Accurately, $T$ represents an instant time if the corresponding behavioral pattern represents a state transition of $o$, or $T$ represents a time interval if the behavioral pattern represents a sequence of state transitions or a state of an attribute. $start$ and $end$ represent the start time and termination time of the geographic object. The function $ex_{aspect}$ that extracts a set of aspects indicated by a behavioral pattern $bpt$ is defined as follows.

$$\begin{aligned} ex_{aspect} &: BehavioralPattern- > \{Aspect\}, \\ ex_{aspect}(bpt) &= \{(obj, T) | obj \text{ behaves like } bpt \text{ at } t \in T.\}. \end{aligned} \tag{6}$$

Here, $obj$ is a geographic object which has a behavior like $bpt$ during the time interval represented by $T$. Concerning to the concept of aspect, functions $period$ and $entity$ which return the time interval and the geographic object respectively, are introduced. When $as = (obj, T)$ is defined, $period$ and $entity$ are represented as follows.

$$period : Aspect- > \{Time\}, period(as) = T. \tag{7}$$

$$entity : Aspect- > Object, entity(as) = obj. \tag{8}$$

**Participant objects in events.** An event has its spatio-temporal field, and each geographic object has its position during the lifespan. Geographic objects which are placed in an event's field are regarded as participants of the event. The notation of event's participant is as follows.

$$\begin{aligned} ex_{participant} &: Event- > \{Aspect\}, \\ ex_{participant}(ev) &= \{(obj, T) | t \in T.contain(pos(t, ev), pos(t, obj)), \\ & start(obj) < t < end(obj)\}. \end{aligned} \tag{9}$$

$pos$ is a function which represents the spatial region of a geographic object $obj$ or an event $ev$ (indicated by the second argument) at time $t$ (indicated by the first argument). $start$ and $end$ represent the onset time and the end time of a geographic object or an event (indicated by an argument), respectively. Consequently, the function $ex_{participant}$ returns a set of geographic objects with valid time intervals while the set of geographic objects exists in the event field.

**Set operation on aspect set.** Participant geographic objects for an event and geographic objects with a similar behavior can be represented as a set of aspects by functions $ex_{participant}$ and $ex_{aspect}$, respectively. Therefore, we can perform the set operation such as $union$, $intersection$, etc. among these sets of aspects. However, general set operations cannot reflect the user intentions properly, since these operations are regardless of the spatio-temporal relation between the elements in two sets of aspects. In this section, we introduce the three constraints for the set operations: spatial constraints, temporal constraints and spatio-temporal constraints.

The spatial constraint is represented by spatial relations between the elements in the sets. The topological relation and distant relation are employed as the spatial relation. The temporal constraint is represented by temporal relations between time intervals, while geographic objects in two sets have common features: i.e. having a common behavior or participating in an event. The temporal relation is based on Allen's interval logic [8]. There are six relations in Allen's interval logic: $before$, $meets$, $overlaps$, $starts$, $contains$, and $finishes$. The spatio-temporal constraint is a combination of spatial relation and temporal relation.

Spatial, temporal, and spatio-temporal constraints restrict a couple of elements in two sets of aspects. The set operations among aspects with these constraints are introduced as follows. Here, the constraints are alternated as $constraint$.

$$\begin{gathered} union(AS_1, AS_2) = \{as | as \in AS_1 \cup AS_2\}, \\ intersection_c(AS_1, AS_2, constraint) = \\ \{as | as \in AS_1, as_2 \in AS_2.constraint(as, as_2) \wedge as = as_2\}, \\ production_c(AS_1, AS_2, constraint) = \\ \{(as_1, as_2) | as_1 \in AS_1, as_2 \in AS_2.constraint(as_1, as_2)\}, \\ selection_c(AS_1, AS_2, constraint) = \\ \{as | as \in AS_1, as_2 \in AS_2.constraint(as, as_2)\}. \end{gathered} \tag{10}$$

$AS_1$ and $AS_2$ are aspect sets of participant geographic objects or geographic objects that have a common behavior. The notation $union(AS_1, AS_2)$ is a general union set of $AS_1$ and $AS_2$. The notation $intersection_c(AS_1, AS_2, constraint)$ is a set of common elements of $AS_1$ and $AS_2$. Moreover, the elements must satisfy $constraint$.

$production_c(AS_1, AS_2, constraint)$ represents a set of couples in $AS_1 \times AS_2$ that satisfy $constraint$. $selection_c(AS_1, AS_2, constraint)$ extracts the elements from $AS_1$ that satisfy $constraint$ for elements in $AS_2$.

In the 2nd layer, the functions, $ex_{aspect}, ex_{participant}$ and set operations are stored, and specified associations are derived as aspect sets from geographic objects and events by using the procedures.

### 2.3 3rd Layer: Event Representation Layer

In order to handle event information in application systems, several notation and operation for representing event feature would be needed. Therefore, in this section, several operations are introduced: aggregation for deriving a value from aspect set, representation field values and extraction of subsets from aspect sets. In the 3rd layer, in order to represent event feature, these procedures are stored.

**Derivation of a value from aspect set.** Aggregate operations derive a value from an aspect set or numerical values of elements of an aspect set. The examples of aggregate operation are the count operation for a set, the average operation for numerical values of aspects and so on. These aggregate operation $f_{agg}$ takes an aspect set and attribute name optionally as arguments, and returns a value.

$$f_{agg} : \{Aspect\}, |attname, t| - > value, f_{agg} \in \{avg, count, max, \ldots\}. \tag{11}$$

$attname$ and $t$ are optional arguments, and they mean the attribute name and the instant time at which an aggregated value is computed, respectively. In addition, all of elements in the aspect set need to have the attribute $attname$ with the same range.

**Field representation.** A field is represented as a couple of a spatial region and a value. A spatial region could be computed from a set of aspects such as a convex hull, a median point and a corresponding administrative district. The function $space$ that computes spatial regions from an aspect set is provided as follows.

$$\begin{gathered}
space : \{Aspect\} - > Region, space \in \{hull, map, center\}, \\
hull(AS) = MIN_{\forall as \in AS, t \in period(as).contains(region, pos(as,t))}(region), \\
map(AS) = MIN_{region \in \Sigma_{s,i} \wedge \forall as \in AS.contains(region, pos(as,t))}(region), \\
center(AS) = median(\cup_{as \in AS}(pos(as,t) | t \in period(as))).
\end{gathered} \tag{12}$$

$hull$ is a function for computing a convex hull, $map$ is for computing a corresponding administrative district, and $center$ is for computing a median point. $MIN_{cond}(region)$ represents a minimum region under a condition $cond$. $\Sigma_{s,i}$ is a set of spatial units such as administrative districts or spatial grid cells. $median$ compute a median point of spatial regions. Consequently, given an aspect set $AS$ ,the fields could be represented as follows.

$$\begin{gathered}
(f_{agg}(AS), space(AS)), \\
f_{agg} \in \{avg, max, min, count, ...\}, space \in \{hull, map, center\}.
\end{gathered} \tag{13}$$

**Extraction subsets from aspect set.** subsets of aspect sets could be obtained using equivalence partitioning among aspects. There are two equivalence relations among aspects: attribute values and spatial units. The attribute partition $div_{atr}$ divides an aspect set into subsets with equivalent values. The spatial partition $div_{space}$ divides an aspect set into subsets of which elements exist in the same spatial unit such as country, prefecture, city or grid cells. In addition, temporal partition $div_{time}$ could be performed. Temporal partition is performed in two steps. First, each aspect is divided into several aspects based on temporal units such as day, week and so on. Second, these divided aspect sets are re-grouped by their periods. A detailed explanation is described in [5].

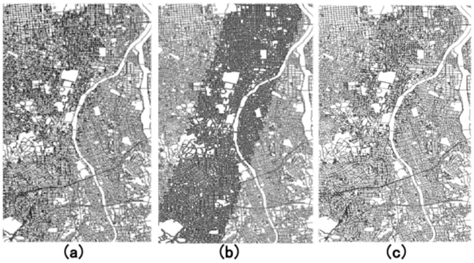


**Fig. 2.** Computation of association between an event and geographic objects

## 3 Computational Examples

Computational examples described in the Section 2 are shown in Fig.2 and Fig.3. Fig.2(a) describes the aspect set of collapse building and road objects(Eq.4); results of $ex_{aspect}(bpt_{collapse})$ are displayed. These are caused by typhoons or other factors. In Fig.2(b), a set of geographic objects that participate in a typhoon event($ev_1$) is visualized. Namely, $ex_{participant}(ev_1)$ is displayed. Fig.2(c) shows the result of the set operation computing intersections of two aspect sets in Fig.2(a) and Fig.2(b), with the time constraint that the relation between two time intervals, participating in the event and behaving like $bpt_{collapse}$, is $overlaps$. This computation is represented formally as $intersection_c(ex_{aspect}(bpt_{collapse}), ex_{participant}(ev_1), overlaps)$. Consequently, the result described in Fig.2(c) could be treated as the roads and buildings collapsed by the typhoon $ev_1$.

In Fig.3, examples of event feature representation from aspect sets are provided. Fig.3(a-c) are parts of the results that are obtained by temporal partition for the aspect set in Fig.2(c) with 12 hours and computing median points for each set. Namely, given the aspect set in Fig.2(c) as $AS$, Fig.3(a-c) are the parts of computational result of $\{Center(SubAS_i)|SubAS_i \in div_{time}(AS)\}$ and they represent trajectories of the typhoon. Fig.3(d) is an example of field representation, which is obtained through three step. First, the aspect set in Fig.2(c) is spatial partition with cities. Then, each aspect set is counted up, and mapped to a corresponding city in the field representation. Lastly, regions in each city are displayed with colors corresponding to the number count. Fig.3(d) shows fields $\{(count(SubAS_i), map(SubAS_i))|SubAS_i \in div_{space}(AS)\}$.

## 4 Consideration

In the proposed structure, the processes of storing events and geographic objects, associating event and geographic objects and computing event features are separated. Therefore, procedures for deriving event information from the distributed databases could be shared among a number of application systems. For example, aspect sets in Fig.2 could be referred from not only a typhoon information system and a disaster information system but also a road network management system, in which the aspect set can be used to specify reasons for changing road states such as “destructed by the typhoon”.

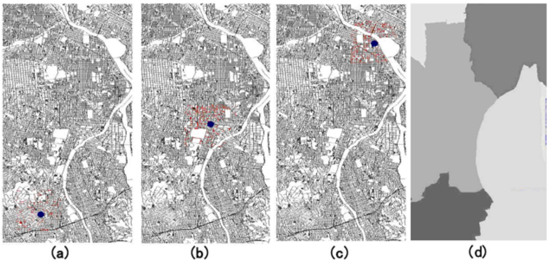


**Fig. 3.** Event feature representation

## 5 Conclusion

In this paper, we proposed the three-layered structure model for sharing and re-using event information in the distributed environment. In our model, event data and geographic objects are represented independently in the 1st layer, and they are associated dynamically according to a query in the 2nd layer. In the 3rd layer, aspect sets from the 2nd layer are modified to represent event features for using in application systems. This framework could realize sharing event information as a set of aspects in 2nd layer.

In our future work, a concrete mechanism for reaching necessary data and avoiding data duplication in the distributed environment. In addition, for developing a practical system, a lot of schemes are needed such as cache mechanism.

## References

1. Galton, A.: Dynamic collectives and their collective dynamics. In: Cohn, A.G., Mark, D.M. (eds.) COSIT 2005. LNCS, vol. 3693, pp. 300–315. Springer, Heidelberg (2005)
2. Worboys, M.F., Hornsby, K.: From objects to events: Gem, the geospatial event model. In: Egenhofer, M.J., Freksa, C., Miller, H.J. (eds.) GIScience 2004. LNCS, vol. 3234, pp. 327–343. Springer, Heidelberg (2004)
3. Goodchild, M.F., Fu, P., Rich, P.: Sharing geographic information: An assessment of the geospatial one-stop. Annals of the Association of American Geographers 97(17), 250–266 (2007)
4. McGregor, S.J.: An integrated geographic information system approach for modeling the suitability of conifer habitat in an alpine environment. Geomorphology 21(3-4), 265–280 (1998)
5. Ikezaki, M., Mukai, N., Watanabe, T.: Event handling mechanism for retrieving spatio-temporal changes at various detailed level. In: Ali, M., Esposito, F. (eds.) IEA/AIE 2005. LNCS (LNAI), vol. 3533, pp. 353–356. Springer, Heidelberg (2005)
6. Ikezaki, M., Ushiama, T., Watanabe, T.: Geographical information structure for managing a set of objects as an event. In: Gabrys, B., Howlett, R.J., Jain, L.C. (eds.) KES 2006. LNCS (LNAI), vol. 4252, pp. 1232–1239. Springer, Heidelberg (2006)
7. Ikezaki, M., Ushiama, T., Watanabe, T.: A geographic event management, based on set operation among geographic objects. In: Apolloni, B., Howlett, R.J., Jain, L. (eds.) KES 2007, Part I. LNCS (LNAI), vol. 4692, pp. 712–719. Springer, Heidelberg (2007)
8. Allen, J.F.: Towards a general theory of action and time. Artif. Intell. 23(2), 123–154 (1984)

# An Introduction to Authorization Conflict Problem in RDF Access Control

Jaehoon Kim, Kangsoo Jung, and Seog Park

Department of Computer Science, Sogang University,
1-1 Shinsu-Dong Mapo-Gu, Seoul, Korea, 121-742
{chris3,jung,spark}@dblab.sogang.ac.kr

**Abstract.** In this paper, related with RDF security, we introduce an RDF triple based access control model considering explicit and implicit authorization propagation. Since RDF Schema represents ontology hierarchy of upper and lower classes or properties, our access control model supports the explicit authorization propagation where an authorization specified against an upper concept is propagated to lower concepts by inheritance. In addition, we consider the implicit authorization propagation where an authorization specified against an lower concept is propagated to upper concepts by RDF inference. RDF Semantics, which is recommended by W3C, guides some primary RDF inference rules related with subClassOf and subPropertyOf where lower concepts are interpreted into upper concepts. Based on these two contrary propagations, we introduce an authorization conflict problem in RDF access control.

**Keywords:** knowledge security, Semantic Web, RDF data, access control, authorization conflict.

## 1 Introduction

RDF and OWL are the primary base technologies for implementing Semantic Web defined by W3C. Considering the access control for RDF data, we can consider simply using the existing XML access control models [1, 2, 3]. This is because RDF and OWL models are described in XML (eXtensible Markup Language). However, as mentioned in several recent studies on the RDF access control model [4, 5], the simple approach is not desirable because XML is used only as the base language for describing RDF model. That is, the most important reason is from ontology inference. When a set of XML access authorizations are explicitly specified for RDF tags, they do not define which authorization should be applied to information that is newly inferred by the ontology inference. Recently, Jain and Farkas [4] have introduced an access control model based on the RDF triple, which considers the inference feature of RDF. We think that their study has more significant meaning compared to several existing studies for RDF access control [5, 6, 7]. Because they represent a security object as an RDF triple that is the basic structure of RDF model and based on the RDF triple, they deal with the problem of authorization conflict in the RDF inference.

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 583–592, 2008.

Here, the authorization conflict problem by ontology inference is as follows. In Fig. 1, let us consider that an access to *NuclearWeapn* is allowed to a user Dave. If so, Dave can browse *Titan*, which is an instance of *NuclearWeapon*. However, let us assume that there is a formerly specified authorization disallowing an access to *SpecialWeapon*. In this case, the access to *Titan* must be canceled because the instance *Titan* can be interpreted as the upper class *SpecialWeapon*.

However, Jain and Farkas did not consider explicit authorization propagation over the ontology hierarchy of upper and lower classes or properties. The explicit propagation of RDF authorizations is that when an authorization is specified for an upper concept, the same access authorization is also applied to all lower concepts over the ontology hierarchy by inheritance. This explicit authorization propagation based on the ontology hierarchy is necessary for a more convenient authorization specification of the security administrator. That is, a variety of authorization specifications can be performed at one time without the need to be performed separately, the number of authorizations can be reduced, and a security administrator can clearly understand the scope of authorization. In this paper, we introduce an RDF triple based access control model supporting explicit propagation as well as implicit propagation. In the suggested method, we first define an RDF authorization specification with the explicit authorization propagation to the lower classes or properties. Then considering the RDF inference, we define implicit authorization propagation to the opposite direction, that is, to the upper classes or properties. Next, using these two contrary propagations, we explain the problem of RDF authorization conflict in RDF access control. We call our suggested access control model RDFacl.

The remainder of the paper is organized as follows. In Section 2, we review several recent studies related to the RDF access control. Next, Section 3 briefly explains the basic characteristics of RDF and RDF Schema related to our study, and Section 4 introduces the proposed RDF authorization model. Section 5 introduces the explicit authorization propagation, and Section 6 introduces the implicit authorization propagation along with the authorization conflict problem. Section 7 finally concludes this paper.

## 2 Related Work

Damiani et al. [1] and E. Bertino et al. [2, 3] suggested the fine-grained access control models for XML documents. According to a specified access authorization, element tags and attributes in an XML tree structure are made to be selectively invisible to users. Even though RDF documents are described in XML, the existing XML access control models do not consider the security violation by ontology inference as mentioned in Section 1.

Qin and Atluri [5] considers the implicit authorization propagation and authorization conflict problem for more various semantic relations in an ontology.

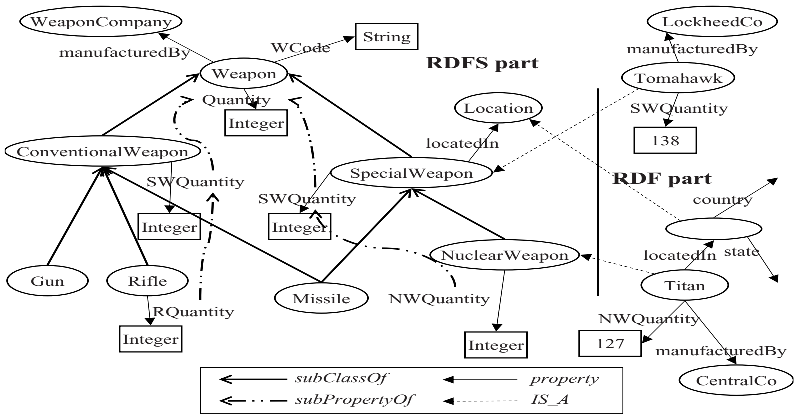


**Fig. 1.** A sample Weapon RDF(S) graph

That is, besides the *subClassOf* and *subPropertyOf* relationships, they consider the equivalence relationship between two concepts, the partial relationship between whole concepts and partial concepts (for example, *intersectionOf* and *unionOf* in OWL), the non-inferable relationship (for example, *disjointWith* and *differentFrom* in OWL), etc. However, their access control policy is not based on the RDF triple structure. Hence, their methods are not incorporated with RDF and OWL specification, and especially RDF inference (named RDF entailment [8]). The semantics of authorization conflict problem explained by them is also different from ours. It is our future work to expand our RDF triple based RDFacl model such that it can also be applied to OWL.

Kaushik et al. [6] introduces an access control model for the fine-grained information disclosure of an RDF web document as in this study. The main point of their study is to introduce a formal framework to provide disclosure control over parts of an ontology. In addition, they introduce applying several methods of information hiding to RDF data, e.g., removing a specific subtree in an ontology tree or renaming a disallowed class or property according to an authorization. However they do not consider the disclosure problem for highly sensitive data by a prohibited inference. In fact, this problem is closely connected with the authorization conflict problem, which we will deal with in this paper. This is because such an information disclosure is most likely to arise when two authorizations having conflict relationship are both allowed.

In the study of Reddivari et al. [7], they also introduce an RDF triple based access control model for RDF data in an RDF store. Since the target security objects are data in an RDF store such as the knowledge based system and the relational database system, the suggested model defines access control over various operations like insert (*insertModel* and *insertSet*), remove (*removeModel* and *removeSet*), update, and read (*see*, *use*). In this paper, since we consider the browsing permission for RDF data in an RDF web document, only read operation is considered.

## 3 RDF and RDFS

RDF Schema (RDFS) provides some primitive constructors for constructing simple ontologies. *rdfs:class* and *rdf:property* define ontology concept, *rdfs:subClassOf* and *rdfs:subPropertyOf* define class and property hierarchies, and *rdfs:domain* and *rdfs:range* define domain and range constraints for properties. For example, in Fig. 1, *NuclearWeapon* is the subclass of *SpecialWeapon*, and *SpecialWeapon* is the subclass of *Weapon*. *NWQuantity* is the subproperty of *SWQuantity*, and *SWQuantity* is the subproperty of *Quantity*.

An RDF document consists of RDF statements, which use ontology concepts defined in an RDFS. An RDF statement is represented as a triple of [$s$, $p$, $o$], and an RDF document can be represented as a graph consisting of RDF triples. An RDFS statement can also be represented as a triple.

**Definition 1 (RDF(S) graph and triple).** *An RDF(S) (= RDF and RDFS) graph is a set of RDF(S) triples. An RDF(S) triple is represented as [s, p, o], where* $s \in SUBJECT$, $p \in PREDICATE$, *and* $o \in OBJECT$.

*- The set SUBJECT includes URI (Uniform Resource Identifier) nodes defining classes or properties in RDFS and instances in RDF, and blank nodes.*

*- The set PREDICATE includes URI nodes referencing properties in RDFS.*

*- The set OBJECT includes the URI nodes of the other classes and instances related by p, blank nodes, and literals.*

For example, in Fig. 1, the RDF triple [*Weapon*, *manufacturedBy*, *WeaponCompany*] has the class URI constant *Weapon* as $s$, the property URI constant *manufacturedBy* as $p$, and the class URI constant *WeaponCompany* as $o$. [*Titan*, *NWQuantity*, 127] has the instance URI constant *Titan* as $s$ and the literal 127 as $o$. [*Titan*, *locatedIn*, _] has a blank node as $o$. The RDF graph in Fig. 1 represents the sample RDF document of Fig. 2.

**Definition 2 (*subClassOf* and *subPropertyOf*).** *If a class* $c_i$ *is the subclass of another class* $c_j$ ($c_i \subset c_j$), $c_i$ *and its instances inherit the properties of* $c_j$, *and* $c_i$ *can be interpreted as* $c_j$ *by inference. If a property* $p_i$ *is the subproperty of another property* $p_j$ ($p_i \subset p_j$), $p_i$ *can be interpreted as* $p_j$ *by inference.*

For example, in Fig. 2, <*ex:NuclearWeapon rdf:ID* =“*Titan*”> can be interpreted as <*ex:SpecialWeapon rdf:ID* =“*Titan*”> and <*ex:NWQuantity rdf:datatype* = “*&xsd:integer*”> can be interpreted as <*ex:SWQuantity rdf: datatype* =“*&xsd:integer*”>.

## 4 RDF Access Authorization

### 4.1 Security Object

In our authorization specification, security objects are RDF triples. A security administrator can conveniently bind up the target RDF triples into the following RDF pattern.

```
<?xml version=“1.0”?>
<rdf:RDF                xmlns:rdf=“http://www.w3.org/1999/02/22-rdf-syntax-ns#”
xmlns:ex=“http://example.org/schemas/weapon#”>
  <ex:NuclearWeapon rdf:ID=“Titan”>
    <ex:manufacturedBy rdf:resource=“ex:CentralCo”/>
    <ex:NWQuantity rdf:datatype=“&xsd;integer”>127</ex:NWQuantity>
    <ex:locatdIn>
      <rdf:Description>
        <ex:country>USA</ex:country> <ex:state>Arizona</ex:state>
      </rdf:Description>
    </ex:locatdIn>
  </ex:NuclearWeapon>
  <ex:SpecialWeapon rdf:ID=“Tomahawk”>
    <ex:manufacturedBy rdf:resource=“ex:LockheedCo”/>
    <ex:SWQuantity rdf:datatype=“&xsd;integer”>138</ex:SWQuantity>
  </ex:SpecialWeapon>
</rdf:RDF>
```

**Fig. 2.** A sample Weapon RDF web document

**Definition 3 (RDF(S) pattern).** *An RDF(S) pattern is represented as an RDF(S) triple [s, p, o], where s and p can be substituted by variables $x and $y, respectively, and o is always the variable $z (In this study, we do not consider much more fine-grained access control according to o values). Also, a blank node for s and o is not allowed.*

For example, in the RDF(S) graph of Fig. 1, for the RDF(S) pattern $pt_1$ = [\$$x$, *NWQuantity*, \$$z$], the matching RDF(S) triples $\mu(pt_1)$ are [*NuclearWeapon*, *NWQuantity*, *literal*] and [*Titan*, *NWQuantity*, *literal*]. Also, $\mu$([*NuclearWeapon*, \$$y$, \$$z$]) = {[*NuclearWeapon*, *manufacturedBy*, *WeaponCompany*], [*NuclearWeapon*, *WCode*, *literal*], [*NuclearWeapon*, *Quantity*, *literal*], [*NuclearWeapon*, *SWQuantity*, *literal*], [*NuclearWeapon*, *locatedIn*, *Location*], [*NuclearWeapon*, *NWQuantity*, *literal*]}, and $\mu$([\$$x$, \$$y$, \$$z$]) matches all edges in the graph.

### 4.2 Access Authorization

The RDF access authorizations are formally defined as follows. This is similar to the authorization type which was used in the XML access control model suggested by Damiani et al. [1].

**Definition 4 (access authorization).** *An access authorization is a five tuple of the form: <subj, obj, act, sign, type>.*

*- subj is the subject to whom the authorization is granted.*

*- obj refers to a security object and in our study, it is an RDF(S) pattern matching RDF(S) triples.*

*- act refers to an action performed against the security object. Since we focus on information disclosure of an RDF web document in this study, only read operation is considered.*

*- sign is (+) if access is allowed, and (−) if access is forbidden.*

*- type is R (= Recursive) if an authorization should be propagated to lower classes and lower properties according to the subClassOf and subPropertyOf relationship, and L (= Local) if an authorization should not be propagated. We will introduce the details of the authorization propagation according to type in Section 5.*

### 4.3 Hidden Portions of an RDF Document According to an Authorization

We consider applying our access control model to information disclosure of RDF documents published over Web. For example, according to the authorization <Dave, [\$*x*, *manufacturedBy*, \$*z*], read, −, R>, the tag <*ex*:*manufacturedBy rdf*:*resource* = "*ex*:*CentralCo*"> in the sample RDF document of Fig. 2 must be invisible to the subject Dave. In this subsection, we define which portions of an RDF document must be hidden according to a specified access authorization. The hidden portions are decided by the type of *o* values ∈ {class or instance URI constant, blank node, literal} in Definition 1. Table 1 summarizes this.

**Example 1.** *According to <Dave, [\$x, manufacturedBy, \$z], read, −, R>, the p = "ex:manufacturedBy" and the o = 'rdf:resource = "ex:CentralCo"' in Fig. 2 are hidden from the instance s = "Titan". However, the actual instance CentralCo referenced by the instance URI constant "ex:CentralCo" is not hidden. According to <Dave, [\$x, locatedIn, \$z], read, −, R>, the p = "ex:locatedIn" and the blank node <rdf:Description> ... </rdf:Description> are hidden. In the case of <Dave, [\$x, NWQuantity, \$z], read, −, R>, the p = "ex:NWQuantity" and the literal 127 are hidden.*

**Example 2.** *According to <Dave, [NuclearWeapon, \$y, \$z], read, −, L>, Dave cannot show all properties of the class NuclearWeapon. If all properties of a class or an instance should be invisible, the whole class or instance*

**Table 1.** Hidden portions according to the type of *o* values

| | URI constant | Blank node | Literal |
|---|---|---|---|
| read (−) | - A property *p* should be hidden from a class or an instance *s*.<br>- Also, the *o* value of an URI constant should be hidden. However, this does not mean that the object referenced by the URI should be hidden. The referencing relationship is only broken. | - *p* should be hidden from *s*.<br>- Also, the *o* value of a blank node should be hidden. | - *p* should be hidden from *s*.<br>- Also, the *o* value of a literal should be hidden. |

*should be invisible, e.g., in Fig. 2, <ex:NuclearWeapon rdf:ID = "Titan"> ... </ex:NulcearWeapon> is hidden.*

## 5 Explicit Propagation by RDF Authorization Specification

When the *type* of an authorization is R, the authorization affects lower classes or properties by inheritance. Prior to introducing implicit propagation in the RDF inference, we define explicit propagation in the RDF authorization specification.

### 5.1 Propagation Policy in *subClassOf* Relationship

As explained in Definition 2, if $c_i \subset c_j$, $c_i$ inherits the properties of $c_j$. Therefore, we define that when an authorization $ca_j$ is specified for a property $p_k$ of $c_j$, $ca_j$ also affects the property $p_k$ of $c_i$. We denote this $subClassOf$ propagation as $ca_j+ \rightarrow ca_i+$ or $ca_j- \rightarrow ca_i-$.

**Example 3.** *When $ca_j$ = <Dave, [SpecialWeapon, SWQuantity, \$z], read, −, R> is specified, $ca_j$ derives <Dave, [NuclearWeapon, SWQuantity, \$z], read, −, R> and <Dave, [Missile, SWQuantity, \$z], read, −, R> by the propagation policy $ca_j- \rightarrow ca_i-$. When $ca_j$ = <Dave, [SpecialWeapon, \$y, \$z], read, −, R> is specified, due to $p(ca_j)$ = \$y, $ca_j$ is also applied to lower classes inheriting all properties of SpecialWeapon. That is, $ca_j$ derives the following authorizations: <Dave, [NuclearWeapon, SWQuantity, \$z], read, −, R>, <Dave, [Missile, SWQuantity, \$z], read, −, R>, <Dave, [NuclearWeapon, locatedIn, \$z], read, −, R>, and <Dave, [Missile, locatedIn, \$z], read, −, R>.*

**Example 4.** *$ca_j$ = <Dave, [SpecialWeapon, *, *], read, −, R> disallows accessing all their own properties of the lower classes as well as all properties inherited from SpecialWeapon. That is, $ca_j$ also derives <Dave, [NuclearWeapon, \$y, \$z], read, −, R>, and <Dave, [Missile, \$y, \$z], read, −, R>.*

**Definition 5 (RDF * pattern).** *This pattern is represented as [s, *, *] as in the above example. This is a special RDF pattern reserved for matching all properties of s's lower classes as well as s.*

### 5.2 Propagation Policy in *subPropertyOf* Relationship

We define that if $p_i \subset p_j$, an authorization $pa_j$ for $p_j$ also affects the property $p_i$. We denote this $subPropertyOf$ propagation as $pa_j+ \rightarrow pa_i+$ or $pa_j- \rightarrow pa_i-$.

**Example 5.** *When $pa_j$ = <Dave, [ConventionalWeapon, CWQuantity, \$z], read, −, R> is specified, $pa_j$ also derives <Dave, [Rifle, RQuantity, \$z], read, −, R>. In the case of $pa_j$ = <Dave, [SpecialWeapon, \$y, \$z], read, −, R>, $pa_j$ also affects all subproperties of all properties of SpecialWeapon. In the case of an RDF * pattern <Dave, [SpecialWeapon, *, *], read, −, R>, $pa_j$ also affects all subproperties of all properties of all subclasses of SpecialWeapon.*

### 5.3 Propagation Policy in IS_A Relationship

An instance $i_l$ of a class $c_i$ follows the authorization $ca_i$. Also, an instance $i_l$ having a property $p_k$ follows the authorization $pa_k$ for $p_k$.

## 6 Implicit Propagation by RDF Inference and Authorization Conflict

In this study, we focus especially on the subsumption relationship, which is related with the key inference in the RDF entailment; $subClassOf$ and $subPropertyOf$ relationship.

### 6.1 Propagation Policy in *subClassOf* Inference

The RDF authorization conflicts can be classified into two types. One is the explicit authorization conflict and another is the implicit authorization conflict. The explicit authorization conflict addresses that there are several authorizations having different *sign* values for the same security object. On the contrary, the implicit authorization conflict addresses that although there are no explicit authorization conflict, a conflict can occur due to ontology inference. Since the explicit authorization conflict is a trivial problem, we concentrate on the implicit authorization conflict in this paper.

• $c_i \subset c_j$, $ca_j+ \rightarrow ca_i+$, $ca_i+' \Rightarrow ca_j+'$ ($conflict$-$free$): As in Fig. 3(a), let us consider $ca_j+ \rightarrow ca_i+$. Then suppose that an authorization $ca_i+'$ for $pt_i$ = [$c_i$, $p_k$, $\$z$] is re-specified afterwards. Since $pt_i$ can be interpreted as $pt_j$ = [$c_j$, $p_k$, $\$z$] by $subClassOf$ inference, the implicit propagation to the upper class $ca_i+'$ $\Rightarrow ca_j+'$ must be considered. In this case, since $sign(ca_j+') \equiv sign(ca_j+)$, there is no conflict.

**Example 6.** *$ca_j$ = <Dave, [SpecialWeapon, locatedIn, $z], read, +, R> drives $ca_i$ = <Dave, [NuclearWeapon, locatedIn, $z], read, +, R> by the explicit propagation. Then suppose that the authorization $ca_i+'$ = <Dave, [NuclearWeapon, locatedIn, $z], read, +, R> is re-specified. Since [SpecialWeapon, locatedIn, $z] can be inferred from [NuclearWeapon, locatedIn, $z], $ca_i+'$ must be also applied to [SpecialWeapon, locatedIn, $z]. That is, $ca_i+'$ drives the implicit propagation $ca_j+'$ = <Dave, [SpecialWeapon, locatedIn, $z], read, +, L>. Since sign($ca_j+'$) $\equiv$ sign($ca_j+$), this is conflict-free.*

• $c_i \subset c_j$, $ca_j+ \rightarrow ca_i+$, $ca_i-' \not\Rightarrow ca_j-'$ ($conflict$-$free$) : In the same manner, let us consider $ca_j+ \rightarrow ca_i+$ and a new authorization $ca_i-'$ specified afterwards for $pt_i$ = [$c_i$, $p_k$, $\$z$]. In this case, since $ca_i-'$ disallows accessing $pt_i$, any related inference cannot occur. Hence, this case is conflict-free.

• $c_i \subset c_j$, $ca_j- \rightarrow ca_i-$, $ca_i-' \not\Rightarrow ca_j-'$ ($conflict$-$free$) : As in Fig. 3(c), let us consider $ca_j- \rightarrow ca_i-$ and a new authorization $ca_i-'$ specified afterwards for

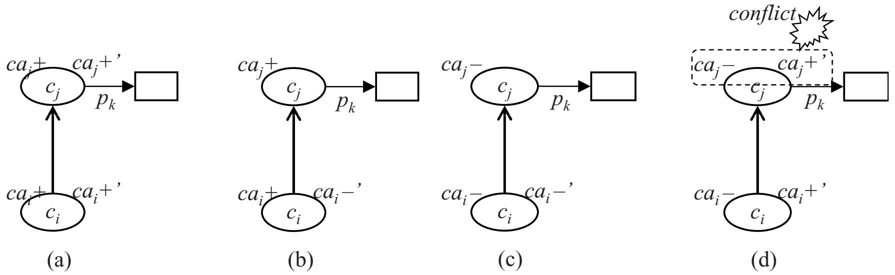


**Fig. 3.** Authorization conflict in *subClassOf* inference

$pt_i = [c_i,\ p_k,\ \$z]$. As in the previous case, since $ca_i-'$ disallows accessing $pt_i$, there can be no conflict.

• $c_i \subset c_j,\ ca_j- \rightarrow ca_i-,\ ca_i+' \Rightarrow ca_j+'\ (conflict)$ : As in Fig. 3(d), let us consider $ca_j- \rightarrow ca_i-$ and a new authorization $ca_i+'$ specified afterwards for $pt_i = [c_i,\ p_k,\ \$z]$. Since $pt_j = [c_j,\ p_k,\ \$z]$ can be inferred from $pt_i$, the implicit propagation $ca_i+' \Rightarrow ca_j+'$ must be considered. In this case, since $sign(ca_j+')$ $\neq sign(ca_j-)$, an authorization conflict occurs. Similarly, when $c_i \subset c_j$, $ca_j+ \rightarrow$ $ca_i+$, and a new authorization $ca_j-$ with $type$ = L is specified afterwards, there is also a conflict.

### 6.2 Propagation Policy in *subPropertyOf* Inference

When $p_i \subset p_j$, $pa_j- \rightarrow pa_i-$, and $pa_i+' \Rightarrow pa_j+'$ as in the *subClassOf* inference, there is a conflict. Figure 4 depicts this situation.

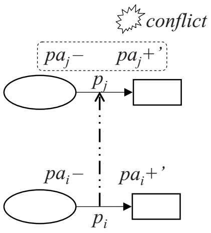


**Fig. 4.** Authorization conflict in *subPropertyOf* inference

## 7 Conclusions and Future Work

The RDF authorization conflict problem should be an important problem in RDF access control because RDF data are related with an ontology inference unlike XML data. Also supporting the explicit authorization propagation in an authorization specification model should be an important problem because specifying an authorization based on ontology hierarchy gives the benefit of more convenient policy description to a security administrator. Therefore, in this paper, we have introduced an RDF access control model considering both the explicit and the implicit authorization propagation. When an ancestor authorization with *sign* = '$-$' and *type* = 'R' is first specified and a descendant authorization with *sign* = '+' is specified afterwards, both authorizations can fall into a conflict relationship. Also, when a descendant authorization with *sign* = '+' is first specified and an ancestor authorization with *sign* = '$-$' and *type* = 'L' is specified afterwards, they can fall into a conflict relationship as well. Currently, based on this observation, we have developed an efficient conflict detection algorithm and are implementing it as a tool. However, due to page limit, we do not present details of that. The basic idea is to check only the ancestor authorizations with *sign* ($-$) when a new authorization is specified with *sign* (+) whereas to check only the descendant authorizations with *sign* (+) when a new authorization is specified with *sign* ($-$) and *type* L.

**Acknowledgments.** This work was supported by the Korea Science and Engineering Foundation(KOSEF) grant funded by the Korea government(MOST) (No. R01-2006-000-10609-0).

## References

1. Damiani, E., Vimercati, S.D.C., Paraboschi, S., Samarati, P.: A fine-grained access control system for XML documents. ACM Transactions on Information and System Security 5(2), 169–202 (2002)
2. Bertino, E., Castano, S., Ferrari, E., Mesiti, M.: Specifying and enforcing access control policies for XML document sources. World Wide Web Journal 3(3), 139–151 (2000)
3. Bertino, E., Ferrari, E.: Secure and selective dissemination of XML documents. ACM Transactions on Information and System Security 5(3), 290–331 (2002)
4. Jain, A., Farkas, C.: Secure resource description framework: an access control model. In: 11th ACM Symposium on Access Control Models and Technologies, pp. 121–129 (2006)
5. Qin, L., Atluri, V.: Concept-level Access Control for the Semantic Web. In: ACM Workshop on XML Security 2003, pp. 94–103 (2003)
6. Kaushik, S., Wijesekera, D., Ammann, P.: Policy-based dissemination of partial web-ontologies. In: The 2005 workshop on Secure web services, pp. 43–52 (2005)
7. Reddivari, P., Finin, T., Joshi, A.: Policy-Based Access Control for an RDF Store. In: The Policy Management for the Web Workshop, pp. 78–83 (2005)
8. RDF Semantics, W3C Recommendation, http://www.w3.org/TR/rdf-mt/

# Improving Text Summarization Using Noun Retrieval Techniques

Christos Bouras and Vassilis Tsogkas

Research Academic Computer Technology Institute, N. Kazantzaki, Panepistimioupoli and Computer Engineering and Informatics Department, University of Patras, Greece
{bouras,tsogkas}@cti.gr

**Abstract.** Text Summarization and categorization have always been two of the most demanding information retrieval tasks. Deploying a generalized, multifunctional mechanism that produces good results for both of the aforementioned tasks seems to be a panacea for most of the text-based, information retrieval needs. In this paper, we present the keyword extraction techniques, exploring the effects that part of speech tagging has on the summarization procedure of an existing system.

**Keywords:** Focused Crawler, Part of Speech Tagging, Noun Retrieval, Data Preprocessing, Text Summarization, Text Categorization.

## 1 Introduction

Keyword extraction, being the basis of any information retrieval (IR) task, aims to select the appropriate keywords out of a text, accompanying them with a suitable score that depicts their importance. With the term appropriate, we mean the most representative words, as far as the text's overall meaning is concerned. Following the keyword extraction procedure, text summarization and categorization techniques come. In our research, a unified, yet autonomous system is developed, PeRSSonal [1], in which summarization and categorization are the core procedures (as well as personalization of the presented results). This paper studies the improvement of the aforementioned procedures by assisting our keyword extraction mechanism with noun retrieval capabilities.

Presenting to the user summaries matching their needs is a very crucial procedure that can assist information filtering. Even though automatic text summarization dates back to Luhn's work in the 1950's, several researchers continued investigating various approaches to the summarization problem up to nowadays. A summary [2] usually helps readers identify interesting articles or even understand the overall story about an event. Most of the times, the summarization approaches are based on a “text-span level” [3], with sentences being the most common type of text-span having each of them rated according to some criteria (e.g. important keywords, lexical chains, etc.). These techniques transform the original problem to a simpler one: ranking sentences according to their salience or likelihood of being part of a summary, concatenating them at a second stage. Some techniques [4] try to identify special words and phrases in the text, while in [5] the authors compare patterns of relationships between the sentences.

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 593–600, 2008.

Typical classification tasks are deciding to what folder an email message should be directed, on which newsgroup a news article belongs, etc. Several text classification (categorization) approaches have been researched over the years: Naive Bayesian(NB), K-Nearest Neighbor(KNN), and Centroid-based(CB) techniques are some examples. New articles can be categorized to the pre-defined categories using some criteria which vary from one technique to another. Categories can be relatively coarse-grained, i.e. only some basic unrelated to each other, or fine-grained, where many categories, frequently overlapping with each other, are introduced. Linear Least Squares (LLSF) [6], a multivariate regression model that is automatically learned from a training set of documents and their categories, gives good results and is utilized in our work.

Automatic part of speech tagging, is a well known problem that has been addressed by several researchers during the last twenty years. It is a firm belief that when it comes to keyword extraction, the nouns of the text carry most of the sentence meaning. In a sense, extracted nouns should lead to better semantic representation of the text, and hence, improved IR results. Noun extraction, a subtask of POS tagging, is the process of identifying every noun (either proper or common) in an article or a document. In many languages, nouns are used as the most important terms (features) that express a document's meaning in NLP applications such as information retrieval, document categorization, text summarization, information extraction, etc. Various methodologies have been proposed making use of linguistic [7], statistical [8], symbolic learning knowledge [9] or support vector machines [10] and can be categorized to: morphological analysis, or POS tagging based. The former methods try to generate all possible interpretations of a given phrase by implementing a morphological analyzer or a simpler method using lexical dictionaries. It may over-generate or extract inaccurate nouns due to lexical ambiguity and shows a low precision rate. On the other hand, the POS tagging based methods choose the most probable analysis among the results produced by the morphological analyzer. Due to the resolution of the ambiguities, it can obtain relatively accurate results. However, it also suffers from errors not only produced by the POS tagger, but also triggered by the preceding morphological analyzer.

In this paper we present the incorporation of noun retrieval techniques in PeRSSonal, using the SVM method for POS tagging, as part of its keyword extraction algorithms, and we explore, though experimentation, the possible improvements this change has on the mechanism's IR procedures: summarization and categorization.

In the next section the architecture of the proposed mechanism is introduced. In Section 3 the algorithm analysis of the mechanism is presented. Section 4 describes the experimental procedure that took place and its results. Section 5 concludes and outlines the directions of possible future research.

## 2 Architecture

PeRSSonal [1] follows a classic n-tier architectural approach. The system consists of four layers which work autonomously and collaborate through a centralized database. The web interface handles the information flow into the mechanism which is then directed to the interior subsystems. Text preprocessing techniques follow and the results are led to the next level of analysis where core IR techniques are located. Finally the outcomes are presented to the end users though the information presentation subsystem. In the current paper, we extend the text preprocessing subsystem using noun retrieval techniques. The implemented architecture is depicted in Fig. 1 and the modified component is presented in the dashed box.

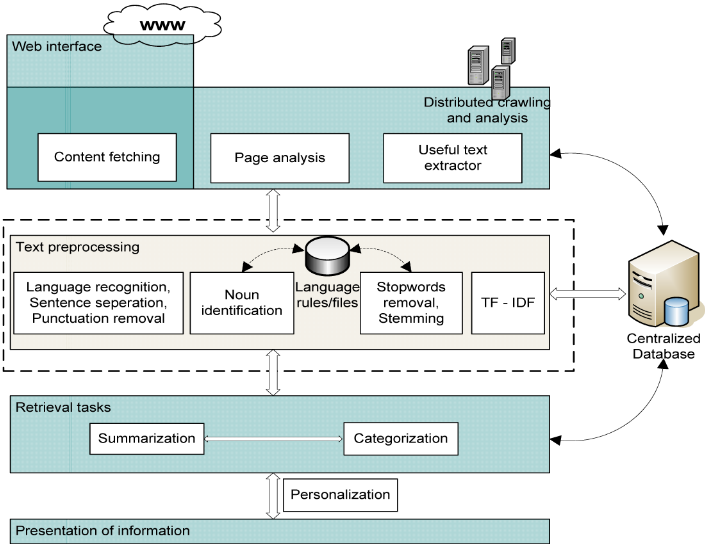


**Fig. 1.** System's architecture

The first layer constitutes the interconnection between the mechanism and the web sources where the following procedures take place: content fetching procedure, analysis of the downloaded content and lastly, extraction of the useful information from the web content. In order to capture web pages, a simple focused web crawler is used. The crawler takes as input the addresses that are extracted from existing RSS feeds, deriving from several major news portals. The crawling procedure is distributed across multiple systems which synchronize thought the centralized database. Crawled html pages are analyzed and are stored without any other unnecessary page element (images, css, javascript, etc.). During this analysis level, our system isolates the "useful text", meaning the main body of the article, and the database is populated with news articles that are ready for the text preprocessing step.

The second tier of the system, which is the focusing of this paper, works on the article's title and body applying several preprocessing techniques. In particular, after the retrieval of the stored article that resides in the database, a series of inner procedures take place at this layer. Firstly, the article's language is recognized either directly through language identifying procedures, or indirectly using the predetermined language of the origin-feed. Following is a sentence separation and punctuation removal step. Afterwards, the noun identification step takes place which, by utilizing the POS SVM-based tagger [10], is able to determine with high precision the article's nouns. Some common text extraction techniques follow: stopwords removal and stemming.

Noun extraction should precede these procedures if it is to succeed will high probability. It is important to note that the noun identification, stopwords removal and stemming procedures are language dependant, meaning that specific language rules, stopword lists and stemming rules respectively, have to be applied for different languages. The above set the foundations for multi-language support by our mechanism, even though only the English language has been incorporated so far. The results of the procedures described in this layer are stemmed keywords either marked as nouns or not, their location in the text and their frequency of appearance in it. These are represented through term frequency – inverse document frequency (TF-IDF) vector statistics that are stored in the database and are utilized by the procedures of the third analysis level.

The information retrieval tasks of our mechanism are located in the third analysis level, where the summarization and categorization algorithms are applied. The main scope of the categorization module is to assist the summarization procedure by prelabeling the article with a category and has proven in [1] to be providing better results. Following the IR task of the mechanism, personalization algorithms take place and the content is finally delivered to the user.

## 3 Algorithm Analysis

Our analysis consists of three different algorithmic steps: extraction of keywords and identification of nouns, categorization procedure and summarization procedure.

### 3.1 Keyword Extraction and Noun Identification Procedure

The input to the keyword extraction module is plain text that defines the article's body and title as well as its language. Apart from the previous, some parameters have to be tuned in order for the mechanism to be the most efficient: a) minimum word length (all words with length smaller than the minimum are removed) and b) the language dependant stopword list that will be used. Our experimentation for news articles in [11] revealed that a limit of 5 letters is the best suited as far as articles written in English are concerned.

Noun identification involves an off-line learning step for the POS tagger using language specific rules. Previous to the tagging, SVM models (weight vectors and biases) are learned from a training corpus using the learning component. A modified version of the SVMTool [10] is used so that tagging takes place only for nouns, saving system's processing time. Once the training is complete, the article's body is forwarded to the tagger and the text's nouns are marked. Stopwords removal takes place and stemming rules are applied, resulting to the TF-IDF vector for all the texts and their terms.

### 3.2 Categorization Procedure

The categorization subsystem is based on the cosine similarity measure, dot products and term weighing calculations. The system is initialized with a training set of 1500 pre-categorized articles, belonging to 7 different categories. The categorization module receives as input the extract of the pre-processing mechanism, which is: a) stemmed keywords, b) noun-related information, c) absolute and relative frequency of the keywords appearance in the article and d) the article's title and body. After the initialization of the training set, the categorization module creates lists of keywords-nouns that are representative of a unique category, consisting of nouns with high frequency at a specific category, and small or zero frequency for the others.

The categorization attempt of a recently fetched article resembles the LLSF method and proceeds as follows; the labeling of the articles is done by using the list of the representative (stemmed) keywords of the text together with the frequencies evaluated by the pre-processing mechanism (Table 1). We then produce identical lists for all the categories that we own that consist of the same keywords followed by their frequency into the category (Table 2). In order to determine the text's category, we examine the cosine similarity of the text and the categories based on the aforementioned lists.

**Table 1.** Article's categorization vector

| Stemmed k/w | Frequency |
|---|---|
| sharia | 4 |
| minist | 7 |

**Table 2.** Politics category vector

| Stemmed k/w | Frequency |
|---|---|
| sharia | 0 |
| minist | 90 |

An article is most of the times related with a similarity measure to more than one category. However, for a categorization result to be accepted we define certain thresholds: (a) the cosine similarity between the text and the category should be over $T_{hr1}$, and additionally (b) the difference of the cosine similarity between the highest ranked category and the rest should exceed $T_{hr2}$. Experimentation, gave us the best suited thresholds for $T_{hr1}$ and $T_{hr2}$ as 0.50 (50% similarity), and 11% respectively. If $T_{hr1}$ or $T_{hr2}$ is not met, the article is forwarded to the summarization module and the resulting generic summary is used as input to a second categorization attempt for the article. Should the above thresholds be met, the labeling of the summary is kept, while at a different case, the initial labeling of the article is kept.

### 3.3 Summarization Procedure

During the summarization procedure, we utilize three factors: (a) the existence of a keyword in the title (b) the frequency of a keyword and (c) the noun tagging information of a keyword. We call these factors k1, k2 and N respectively. A keyword with very high frequency in the text is considered to be representative of it and thus, any sentence that includes it can be considered as text-representative. Additionally, any keyword of the text that also exists in the title is marked as an important one, so the sentences that include it are more representative. Moreover, when a keyword is tagged as a noun, we consider it significant thus boosting it with some extra weight. Parameters k1 and k2 are thoroughly explained in [1]. N derives from the following equation:

$$N = L * z \tag{1}$$

where z=0 if the keyword is not a noun and z=1 if it is. L conveys the desired extra weight that a noun existing in a sentence should have. Experimentation with various L values revealed that L should be no more than 1.5 or else sentences with few keywords-nouns receive low scores, compared to sentences with many nouns, and are substantially excluded from the summary. Typical values for L range from 0 to 1 with the former depicting that the summarization algorithm is not taking into consideration the noun relevant information.

Based on these heuristics, we create a summary which consists of the most representative sentences of the text. In order to determine these, we deploy a score for each sentence according to the factors k1, k2 and N. Assuming that the text T has s sentences where i = [1..s] and f keywords where k = [1..f], each sentence is assigned a score according to the following equation:

$$W_i = \sum (1 + rel(fr(kw_{k,i})))(k_1 + k_2 + N) \tag{2}$$

where rel(fr(kwk,i)) is the relative frequency of the keyword k in sentence i.

After creating a generic summary, we retry to achieve a categorization, as the summarized text is more refined and consists only of important sentences rather than the whole text, which may include sentences with keywords that are distracting the categorization procedure.

The procedure that is followed in order to summarize a text after a successful categorization, differs from the aforementioned steps due to the fact that another factor is included in the scoring. This factor, namely k3 in [1], is the keyword's ability to represent the category to which the document belongs. As long as the text is categorized, we can utilize this factor in order to create a more efficient summary. With the use of k3, the overall weighting equation is depicted below.

$$W_i = \sum (1 + rel(fr(kw_{k,i})))(k_1 + k_2 + N)k_3 \tag{3}$$

## 4 Experimental Procedure and Results

In order to evaluate the summarization performance of PeRSSonal, with the appliance of noun retrieval techniques, we conducted two sets of experiments. Firstly, we tried to determine the best possible value for the L parameter of equation (1). Furthermore, we tried to evaluate the effect of the appliance of the noun retrieval algorithm explained earlier, to the overall system performance using classic IR measures. For conducting the experiments we utilized a corpus of 3000 news articles from various sources. The articles belonged with high relevance to one of the seven major categories of the system, and this information was used as explained at the previous section (precategorized articles) in order for the summarization procedure to produce the best possible summary.

As reported earlier, the parameter L is deployed for controlling the effect that noun retrieval has on the summarization procedure. We conducted experiments tuning L in order to decide on its best value as far as news articles, which is the case of PeRSSonal, are concerned. The various results are presented in Fig. 2.

At the previous graph it is clearly depicted that a value between 0.5 and 0.6 for L is best fitted. Values for L over 1 seem to attenuate both the precision and the recall of the summarization procedure compared to the L=0 case, i.e. when noun retrieval information is not used. This intuitively means that, when sentences that contain mostly

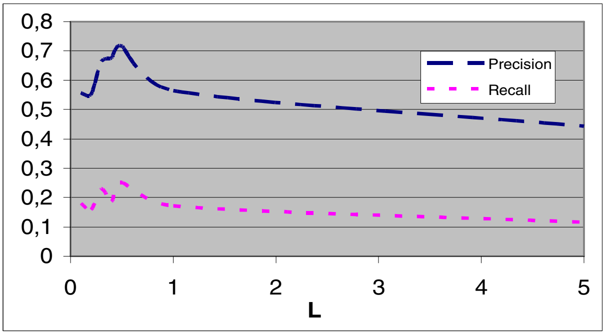


**Fig. 2.** Precision / recall results for summarization of news articles tuning the L value

nouns are kept at the summarization procedure, excluding the rest of the sentences, the effectiveness of the procedure slightly deteriorates. However, finding a golden section for the L parameter, which is dependable on the target texts, can enhance the summarization efficiency significantly. This is also obvious at the following graph where precision and recall results are depicted (using an L value of 0.6) when summarization proceeds with and without the noun-retrieval information.

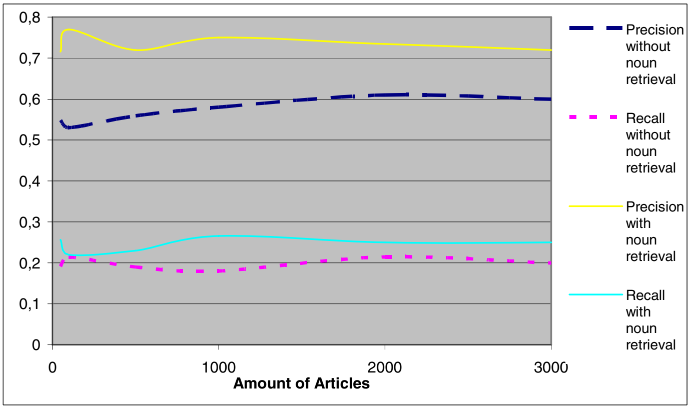


**Fig. 3.** Precision and Recall results for the PeRSSonal's summarization procedure when noun retrieval information is utilized and not

From Fig. 3 it is concluded that the noun retrieval information can give a notable precision boost to the resulting summaries compared to the case where noun retrieval information is not utilized; in other words, the resulting summaries are more precise. As far as recall is concerned, the improvement is small, yet significant, taking into account the fact that a text's summary represents a layer of abstraction, notably a low recall representation of the original text's information.

## 5 Conclusions and Future Work

In this paper we explored the effects that noun retrieval techniques, based on POS tagging, can have on information retrieval mechanisms and summarization in specific. Through the proposed framework that is utilized in an existing system, PeRSSonal, we are able to improve the summarization procedure by simple modifications to our keyword extraction algorithm. The efficiency improvements are small yet significant considering the fact that summarization is a difficult, mostly subjective procedure and that objective criteria of efficiency are difficult to appoint. Having incorporated noun retrieval techniques we are focusing on multilingual and multimedia support for PeRSSonal, the addition of which should require a throughout redesign of the main parts that consist the system. Also, we are considering a wider evaluation of the improvements that the applied noun-retrieval technique has on both the summarization and the categorization procedure.

## References

[1] Bouras, C., Poulopoulos, V., Tsogkas, V.: PeRSSonal's core functionality evaluation: Enhancing text labeling through personalized summaries. Elsevier Science Publishers B. V, vol. 64, pp. 330–345. Elsevier, Amsterdam (2008)

[2] Wasson, M.: Using Leading Text for News Summaries: Evaluation Results and Implications for Commercial Summarization Applications. In: Proceedings of ICCL (1998)

[3] Goldstein, J., et al.: Summarizing Text Documents: Sentence Selection and Evaluation Metrics. In: Proceedings of ACM SIGIR Conference (1999)

[4] Ferragina, P., Gulli, A.: A Personalized Search Engine Based on Web-Snippet Hierarchical Clustering. In: Proceedings of WWW Conference (2005)

[5] Hayes, P.J., et al.: A News Story Categorization System. In: Proceedings of the second Conference on Applied Natural Language Processing (1988)

[6] Yang, Y., Chute, C.G.: An example-based mapping method for text categorization and retrieval. ACM Transaction on Information Systems (TOIS) 12(3), 252–277 (1994)

[7] Karlson, F., Voutilainen, A., Heikkila, J., Anttila, A.: Constraint Grammar. A Language-Independent System for Parsing Unrestricted Text. Mouton de Gruyter, Berlin (1995)

[8] Cutting, D., Kupiec, J., Pedersen, J., Sibun, P.: A practical part-of speech tagger. In: Proceedings of the 3rd Conference on Applied Natural Processing (1992)

[9] Roth, D., Zelenko, D.: Part-of-Speech Tagging Using a Network of Linear Seperators. In: Proceedings of the 36th Annual Meeting of the ACL – Coling, Mondreal, Canada (1998)

[10] Gimenez, J., Marquez, L.: SVMTool: A general POS tagger generator based on Support Vector Machines. In: Proceedings of the 4th International Conference on Language Resources and Evaluation, pp. 43–46 (2004)

[11] Bouras, C., Poulopoulos, V., Tsogkas, V.: The importance of the difference in text types to keyword extraction: Evaluating a mechanism. In: 7th International Conference on Internet Computing (ICOMP 2006), Las Vegas, Nevada, pp. 43–49 (2006)

# Providing Flexible Queries over Web Databases

Xiangfu Meng, Z.M. Ma, and Li Yan

College of Information Science and Engineering, Northeastern University,
110004 Shenyang, China
marxi@126.com, mazongmin@ise.neu.edu.cn, yanlily@yahoo.com

**Abstract.** This paper presents a novel approach of the flexible query and ranking (FQR) by relaxing the original query in order to provide approximate answers to the user. FQR extends the categorical query criteria with the most similar values by estimating the similarity of different pairs of values in the query workload. Also FQR expands the numerical query criteria range to the nearby values by using the kernel density estimation technology. FQR speculates the importance of each specified attribute based on the user query and assigns the score of each attribute value according to its "desirableness" to the user. The tuples satisfying the relaxed query are finally ranked according to their satisfaction degree.

**Keywords:** Web database, Flexible query, Query results ranking.

## 1 Introduction

With the rapid expansion of the World Wide Web, more and more Web databases[1] are available for lay users. Database query processing models have always assumed that the user knows what he or she want and is able to formulate query that accurately express the query intentions. However, users often have insufficient knowledge about database contents and structure, and their query intentions are usually vague or imprecise as well. Therefore, the query user submits cannot act as rigid constraints for the query results. In another words, the query conditions should be flexible constraints for presenting more information that can meet user's needs and preferences closely.

Some researches have been proposed to handle the flexibility of queries in the database systems. These researches can be classified into two main categories. The first one is based on Fuzzy Sets Theory [12]. Tahani firstly advocated the use of fuzzy sets for querying conventional databases [10]. SQLf language proposed by Bosc and Pivert [2] represents a synthesis of the characteristics and functionalities suggested in other previous proposals of flexible query in classical databases, such as [3, 7, 11]. The second category focuses on the development of cooperative database systems such as ARES [4], MULTOS [8], and PREFERENCES [5], which handle the flexibility based on distance notion, linguistic preferences, and etc.

[1] We use the term "Web database" to refer to a non-local autonomous database that is accessible only via a Web (form) based interface.

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 601–606, 2008.

It should be pointed out that the flexibility approaches based on fuzzy sets are highly dependent on the domain knowledge, and it is mainly useful in expanding the numerical query criteria as well. The cooperative database systems are not fully automatic and require the user feedback. Recently, Nambiar [6] presented an approach for query relaxation that used approximate functional dependencies, but it is only focused on relaxing the categorical query criteria. In this paper, we propose a novel relaxation approach FQR (flexible query & ranking), which can relax both categorical and numerical query criteria ranges and does not require user feedback. This approach uses data and query workload statistics in order to assist the relaxation process. Furthermore, for the relevant answers returned by a flexible query, FQR ranks them according to their satisfaction degree.

## 2 Problem Definition

Consider an autonomous Web database $D$ with categorical and numerical attributes $A = \{A_1, A_2,\ldots, A_m\}$ and a selection query $Q$ over $D$ with a conjunctive selection condition of the form "$A_1 = a_1$ AND $A_2 = a_2$...AND $A_s = a_s$". Each $A_i$ in the query condition is an attribute from $A$ and $a_i$ is a value in its domain. The set of attributes $X = \{A_1, \ldots, A_s\} \subseteq A$ is known as the set of attributes specified by the query. When the query leads to empty or unsatisfactory answers, the original query condition should be relaxed to provide additional similar answers for users.

## 3 Query Relaxation

In order to recommend the similar answers to the user, we need to measure the similarity between the different pairs of values. The idea of query relaxation is to expand the original query criteria with similar values.

### 3.1 Relaxation of Categorical Query Conditions

We discuss an approach for deriving the similarity coefficient of categorical attribute values by using query workload-*log of past user queries* on the database. The workload can reflect the frequency with which database attributes and values are often requested by users and thus may be interesting to new users. The intuition is that if certain pairs of values <$u$, $v$> often "occur together" in the workload, they are similar. Let $f(u, v)$ be the frequency of the values $u$ and $v$ of categorical attribute $A$ occurring together in a *IN* clause in the workload. Also let $f(u)$ be the frequency of occurrence of the value $u$ of categorical attribute $A$ in a *IN* clause in the workload, and $f(v)$ be the frequency of occurrence of the value $v$ of categorical attribute $A$ in a *IN* clause in the workload. Then, we measure the similarity coefficient between $u$ and $v$ by using the following Equation (1).

$$Sim(u, v) = \frac{f(u,v)+1}{\max(f(u), f(v))+1} \tag{1}$$

The Equation (1) indicates that, the more frequently occurring together of the same pair of attribute values is, the larger their similarity coefficient is. Note that, in this formula the *Sim*($u$, $v$) would not be zero even if the pair of values is never referenced in the workload. According to the similarity coefficients between different pairs of categorical attribute values, FQR can provide the most similar information for the user when she submits a query with a relaxation threshold.

### 3.2 Relaxation of Numerical Query Conditions

Unlike categorical values, the similarity coefficient of numerical values is difficult to determine because of the continuity of numerical data. We propose an approach, which is inspired by the fuzzy logic [12], to estimate the similarity coefficient between a pair of different numerical values. Let $\{t_1, t_2, \ldots, t_n\}$ be the values of numerical attribute $A$ occurring in the database. Then the similarity coefficient *Sim*($t$, $q$) between $t$ and $q$ can be defined by Equation (2) as follows, where $h$ is the *bandwidth* parameter and is illustrated in the following.

$$Sim(t, q) = \frac{1}{1+\left(\frac{t-q}{h}\right)^2} \tag{2}$$

Let α be a given relaxation threshold, and $q$ be the numerical value specified by the query. Then we can get the expanded range as follows.

$$[q - h\sqrt{\frac{1-a}{a}}\,,\, q + h\sqrt{\frac{1-a}{a}}\,] \tag{3}$$

A popular estimation for the bandwidth is $h = 1.06\sigma n^{-1/5}$, where σ is the standard deviation of $\{t_1, t_2, \ldots, t_n\}$[9] and $n$ is the number of tuples in the database.

As discussed above, with different threshold that the user chooses for the original query, the user's original queries are translated into flexible queries. However, for a large size database, a flexible query may result in too many relevant answer items. So it is necessary to rank the query results in terms of their satisfaction degree.

## 4 Ranking Relevant Answers

### 4.1 Attribute Weight Assignment

In the real world, different users have different preferences. As a result, the importance of the same attribute is usually different for users. Hence, we need to measure the importance of attribute (i.e. attribute weight) for the user. To some extent, the importance of the specified attribute for the user can be reflected by the distribution of its value specified by the query in the database.

**Importance of the Specified Categorical Attribute.** The well-known *IDF* method has been used extensively in IR. The *IDF* suggests that commonly occurring words convey less information about user's needs than rarely occurring words, and thus should be weighted less. *IDF*($w$) of a word $w$ is defined as $\log(n/F(w))$, where $n$ is thenumber of documents and $F(w)$ is the number of document in which $w$ appears. If the database only has categorical attributes, each tuple can be treated as a small document. Thus, we can mimic these techniques for weighting the attribute values.

For a point query "$A_i = t$", we define $IDF_i(t)$ as $\log(n/F_i(t))$ to represent the importance of attribute value $t$ in the database, where $n$ is the number of tuples in the database and $F_i(t)$ is the frequency of tuples in the database of $A_i = t$. In the paper, the importance of categorical attribute value specified by the query is treated as the importance of its corresponding attribute.

**Importance of the Specified Numerical Attribute.** For valuating the importance of numerical attribute values, it is inappropriate to adopt the definition of traditional *IDF* mentioned above because of their binary nature (where if $u$ and $v$ are arbitrarily close to each other yet distinct). Moreover, the "frequency" of a numeric value should depend on nearby values.

According to [1], we present a definition for estimating the importance of numeric attribute values. Let $\{t_1, t_2, \ldots, t_n\}$ be the values of attribute $A$ that occur in the database. For any value $t$, $IDF(t)$ is defined as follows.

$$IDF(t) = \log\left(\frac{n}{\sum_{i}^{n} e^{-\frac{1}{2}\left(\frac{t_i - t}{h}\right)^2}}\right) \tag{4}$$

Here the parameter $h$ is defined as the same as the above. Intuitively, the denominator in Equation (4) represents the sum of "contributions" to $t$ from every the other point $t_\text{i}$ in the database. These contributions are modeled as scaled distributions, so that the further $t$ is from $t_\text{i}$, the smaller is the contribution from $t_\text{i}$. In this paper, the importance of numerical attribute value specified by the query is treated as the importance of its corresponding attribute.

Consequently, the weight of attribute $A_i$ specified by the query can be defined by

$$W(A_i) = \frac{IDF_i(t)}{\sum_{i=1}^{k} IDF_i(t)} \tag{5}$$

## 4.2 Query-Tuple Similarity Estimation

We measure the similarity between a flexible query $Q$ and an answer tuple $T$ as

$$SIM(Q, T) = \sum_{i=1}^{k} W(A_i) \times Sim(Q.A_i, T.A_i) \tag{6}$$

Here $k = Count(boundattributes(Q))$, $W(A_i)$ is the importance weight of attribute $A_i$ specified by the flexible query $Q$, and $Sim(.)$ is the function which measures the similarity coefficients between categorical or numerical attribute values as explained in Section 3. According to the similarity score, the relevant answers can be ranked.

## 5 Experiments

The experiments aim at evaluating the quality of FQR algorithm for ranking the flexible query results. For our experiments, we set up a used car database CarDB (*Make&Model*, *Year*, *Price*, *Location*, *Mileage*) containing 100,000 tuples extracted from Yahoo! Autos. The attributes *Make&Model*, *Year* and *Location* are categorical attributes and the attributes *Price* and *Mileage* are numerical attributes. In addition to FQR described above, we implement two other ranking methods, RANDOM and QFIDF, which are described briefly below, to compare with FQR. In the RANDOM ranking model, the tuples in the query results are presented to the user in a random order. In the QFIDF ranking model, the ranking score is the similarity between tuple $T$ =<$t_1$,…,$t_m$> and query $Q$ = <$C_1$,…,$C_m$>, which is simply the sum of corresponding similarity coefficients over all attributes specified in the query. The QFIDF technique addresses the similar problem (i.e., ranking of the relevance answers) as does FQR. Our approach differs from that in [1] is as follows. When measuring the tuple's ranking score, QFIDF focuses only on the value exactly matches the query on the corresponding attribute and ignores the similarity of different values, while FQR takes the similarity of different values into consideration. We use a standard *collaborative filtering metric R* provided in [1] to measure ranking quality. Fig. 1 shows the ranking precision of the different ranking algorithms for each test query.

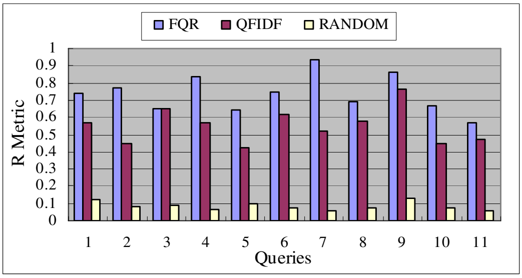


**Fig. 1.** The ranking precision of FQR, QFIDF and RANDOM for each test query

It can be seen that both FQR and QFIDF greatly outperform RANDOM. The averaged ranking precision of FQR and QFIDF were 0.74 and 0.55, respectively. While these preliminary experiments indicate that FQR is promising and better than the existing work, a much larger scale user study is necessary to conclusively establish this finding.

## 6 Conclusions

In this paper, a novel approach for supporting flexible queries over autonomous Web databases is proposed. Starting from the user query, we extend the original query criteria by adding the most similar values into query criteria ranges. For ranking the relevant answers, we assign a weight for each attribute specified by the query according to its importance to the user. Then, for each value of specified attributes in each tuple of the query result, a similarity score is assigned according to its desirableness to the user. All similarity scores are combined according to the attribute weight assigned to each specified attribute. No domain knowledge or user feedback is required in the whole process. Experimental results show that FQR captures user preference fairly well and better than the existing works.

**Acknowledgments.** Work is supported by the Program for New Century Excellent Talents in University (NCET-05-0288) and in part by the MOE Funds for Doctoral Programs (20050145024).

## References

1. Agrawal, S., Chaudhuri, S., Das, G., Gionis, A.: Automated ranking of database query results. ACM Transactions on Database Systems 28(2), 140–174 (2003)
2. Bosc, P., Pivert, O.: SQLf: a relational database language for fuzzy querying. IEEE Transactions on Fuzzy Systems 3(1), 1–17 (1995)
3. Bosc, P., Galibourg, M., Hamon, G.: Fuzzy querying with SQL: extensions and implementation aspects. Fuzzy Sets Systems 28, 333–349 (1988)
4. Ichikawa, T., Hirakawa, M.: ARES: A relational database with the capability of performing flexible interpretation of queries. IEEE Transactions on Software Engineering 12(5), 624–634 (1986)
5. Kießling, W.: Foundations of preferences in database systems. In: International Conference on Very Large Data Bases, pp. 311–322. Morgan Kaufmann Publishers, Hongkong (2002)
6. Nambiar, U., Kambhampati, S.: Answering imprecise queries over web databases. In: 22nd International Conference on Data Engineering, pp. 45–54. IEEE Computer Society, Los Alamitos (2006)
7. Nakajima, H., Sogoh, T., Arao, M.: Fuzzy database language and library: Fuzzy extension to SQL. In: 2nd IEEE International Conference on Fuzzy Systems, pp. 477–482. IEEE Press, San Francisco (1993)
8. Rabitti, F.: Retrieval of multimedia documents by imprecise query specification. In: Bancilhon, F., Tsichritzis, D.C., Thanos, C. (eds.) EDBT 1990. LNCS, vol. 416, pp. 202–218. Springer, Heidelberg (1990)
9. Silverman, B.W.: Density estimation for Statistic and Data Analysis. Chapman and Hall, New York (1986)
10. Tahani, V.: A conceptual framework for fuzzy querying processing: a step toward very intelligent databases systems. Information Processing Management 13, 289–303 (1997)
11. Wong, M., Leung, K.: A fuzzy database-query language. Information Systems 15(5), 583–590 (1990)
12. Zadeh, L.A.: Fuzzy Sets. Information and Control 8(3), 338–356 (1965)

# Fast Cryptographic Privacy Preserving Association Rules Mining on Distributed Homogenous Data Base

Mahmoud Hussein, Ashraf El-Sisi, and Nabil Ismail

CS Deptartment , Faculty of computers and Information, Menofyia University,
Shebin Elkom 32511, Egypt
fci_3mh@yahoo.com, ashrafelsisi@hotmail.com

**Abstract.** Privacy is one of the most important properties of an information system must satisfy. In which systems the need to share information among different, not trusted entities, the protection of sensible information has a relevant role. A relatively new trend shows that classical access control techniques are not sufficient to guarantee privacy when data mining techniques are used in a malicious way. Privacy preserving data mining algorithms have been recently introduced with the aim of preventing the discovery of sensible information. In this paper we propose a modification to privacy preserving association rule mining on distributed homogenous database algorithm. Our algorithm is faster than old one which modified with preserving privacy and accurate results. Modified algorithm is based on a semi-honest model with negligible collision probability. The flexibility to extend to any number of sites without any change in implementation can be achieved. And also any increase doesn't add more time to algorithm because all client sites perform the mining in the same time so the overhead in communication time only. The total bit-communication cost for our algorithm is function in (N) sites.

**Keywords:** association rule mining, apriori, cryptography, distributed data mining, privacy, security.

## 1 Introduction

Privacy preserving data mining is an important property that any mining system must satisfy. There are many methods for privacy preserving distributed association rule mining across private databases. So these methods try to compute the answer to the mining without revealing any additional information about user privacy. An application that needs privacy preserving distributed association rule mining  across private databases, like medical research. There are some existing techniques that one might use for building this application, but they are inadequate related to some disadvantages. One from these techniques is trusted third party. The main parties give the data to a "trusted" third party and have the third party do the computation [1]. However, the third party has to be completely trusted, both with respect to intent and competence against security breaches. The level of trust required is too high for this solution to be acceptable. Also data perturbation technique [2]. Another approach is secure multi-party computation. In this approach given two parties with inputs x and y

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 607–616, 2008.

respectively, the goal of secure multi-party computation is to compute a function f(x,y) such that the two parties learn only f(x,y), and nothing else. In [3] there are various approaches to this problem. In [4] an efficient protocol for Yao's millionaires' problem showed that any multi-party computation can be solved by building a combinatorial circuit, and simulating that circuit. A variant of Yao's protocol is presented in [5] where the oblivious transfers is used to make secure decision tree learning using ID3 with efficient cryptographic protocol and their also two solution of our problem under the secure multi party computation for association rule mining [6], [7]. In this paper we addresses the problem of computing association rules when databases belonging to sites and each site needing preserving the privacy of users data in databases. We assume homogeneous databases: All sites have the same schema, but each site has information on different entities. The goal is to produce a modification to algorithm in [7] that compute association rules that hold globally while limiting the information shared about each site in order to increase the efficiency of the algorithm. The organization of this paper is as follows. Section 2 gives an overview about the problem and the related work in the area of privacy preserving association rule mining on distributed homogenous databases. In section 3 the details of the modification for the algorithm of computing the distributed association rule mining to preserve the privacy of users. Section 4 describes implementation and results of our new algorithm verse the old algorithm. Finally, some conclusions are put forward in Section 5.

## 2 Distributed Association Rule Mining Problem and Related Work

Association Rule mining is one of the most important data mining tools used in many real life applications. It is used to reveal unexpected relationships in the data. We assume homogeneous databases. All sites have the same schema, but each site has information on different entities.

### 2.1 Association Rule Mining

Association rule mining finds interesting associations and/or correlation relationships among large sets of data items. Association rules show attributes value conditions that occur frequently together in a given dataset. A typical and widely-used example of association rule mining is market basket analysis.

In [8] the association rules mining problem can formally be defined as follows: Let $I = \{i_1, i_2, \ldots, i_n\}$ be a set of items. Let DB be a set of transactions, where each transaction T is an itemset such that $T \subseteq I$. Given an itemset $A \subseteq I$, a transaction T contains A if and only if $A \subseteq T$.

An association rule is an implication of the form $A \Rightarrow B$ where $A \subseteq I$, $B \subseteq I$ and $A \cap B = \varnothing$. The rule $A \Rightarrow B$ has support S in the transaction database DB if S% of transactions in DB contains $A \cup B$. The association rule holds in the transaction database DB with confidence C if C% of transactions in DB that contain A also contains B. An itemset X with k items is called a k-itemset.

## 2.2 Distributed Association Rule Mining Problem

The problem of mining association rules is to find all rules whose support and confidence are higher than certain user specified minimum support and confidence. Clearly, computing association rules without disclosing individual transactions is straightforward. We can compute the global support and confidence of an association rule AB ⇒ C knowing only the local supports of AB and ABC, and the size of each database:

$$Support_{AB\Rightarrow C} = \frac{\sum_{i=1}^{numberofsi\ tes} Support\_count_{ABC}(i)}{\sum_{i=1}^{numberofsi\ tes} database\_Size(i)} \qquad Support_{AB} = \frac{\sum_{i=1}^{numberofsi\ tes} Support\_count_{AB}(i)}{\sum_{i=1}^{numberofsi\ tes} database\_Size(i)} \tag{1}$$

$$Confidence_{AB\Rightarrow C} = \frac{Support_{AB\Rightarrow C}}{Support_{AB}} \tag{2}$$

Note that this require no sharing of any individual transactions. What if this information is sensitive? Clearly, such an approach will be secure under secure muti-party computation (SMC) definitions by some modification, a way to convert the above simple distributed method to a secure method in SMC model is to use secure summation and comparison methods to check whether threshold are satisfied for every potential itemset. For example, for every possible candidate 1-itemset, we can use the secure summation and comparison protocol to check whether the threshold is satisfied. Fig. (1) gives an example of testing if itemset ABC is globally supported. Each site first computes its local support for ABC, or specifically the number of itemsets by which its support exceeds the minimum support threshold (which may be negative). The parties then use the secure summation algorithm (the first site adds a random (R) to its local excess support, then passes it to the next site to add its excess support, etc. and finally when pass to first site subtract the generated random from the result). The only change is the final step, the last site performs a secure comparison with the first site to see if the sum ≥ R. In the example, R -10 is passed to the second site, which adds its excess support (5) and passes it to site 3. Site 3 adds its excess support; the resulting value (22) is tested using secure comparison to see if it exceeds the Random value (21). It is, so itemsets ABC is supported globally.

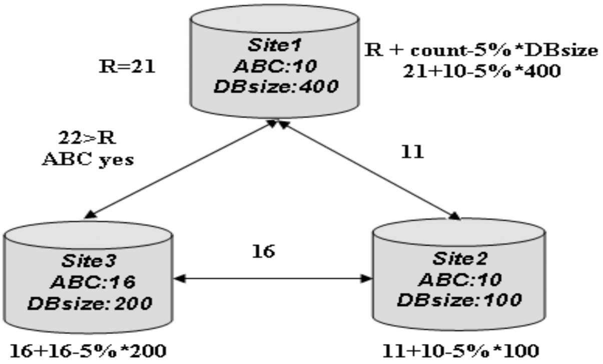


**Fig. 1.** Computing global support securely

Due to huge number of potential candidate itemsets, we need to have a more efficient method. This can be done by observing this lemma, (If a rule has support > k% globally, it must have support > k% on at least one of the individual sites). A distributed algorithm for this would work as follows, request that all rules are sent by each site with support at least k, for each rule returned, request that all sites send the count for their transactions that support the rule, and the total count of all transactions at the site. From this, we can compute the global support of each rule, and be certain that all rules with support at least k have been found. This has been shown to be an effective pruning technique [9]. But the function now being computed reveals more information than the original association rule mining function. However, the key is that we have provable limits on what is disclosed.

### 2.3 Related Work

In [6] an explanation of an efficient method for this problem. To obtain an efficient solution without revealing what each site supports, they instead exchange locally large itemsets in a way that obscures the source of each itemset. They assume a secure commutative encryption algorithm with negligible collision probability. Intuitively, under commutative encryption, the order of encryption does not matter. If a plaintext message is encrypted by two different keys in a different order, it will be mapped to the same cipher text. Formally, commutatively ensures that $E_{k1}(E_{k2}(x)) = E_{k2}(E_{k1}(x))$. The main idea is that each site encrypts the locally supported itemsets, along with enough "fake" itemsets to hide the actual number supported. Each site then encrypts the itemsets from other sites. An example illustrate the protocol in [6] is given in fig. (2). An approach to proof that protocol preserves privacy can be found in [3]. This approach to prove that algorithm reveals only the union of locally large itemsets and a clearly bounded set of innocuous information.

Other method in [7] showed that the protocol in [6] employs commutative encryption algorithm so it adds large overhead to the mining process then another protocol improves this by applying a public-key cryptosystem algorithm on horizontally partitioned data among three or more parties. In this protocol, the parties can share the union of their data without the need for an outside trusted party. Each party works locally finding all local frequent itemsets of all sizes. Then use public key cryptography to find the

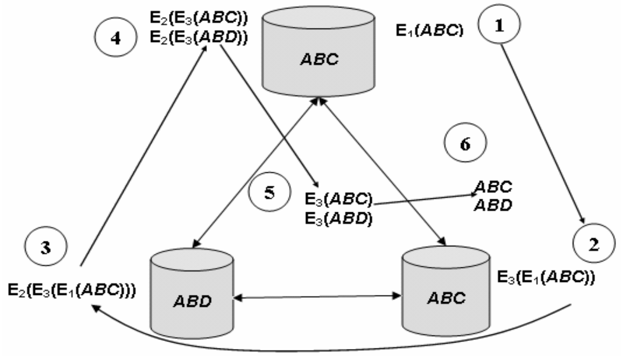


**Fig. 2.** Steps needed for computing the algorithm in [3]

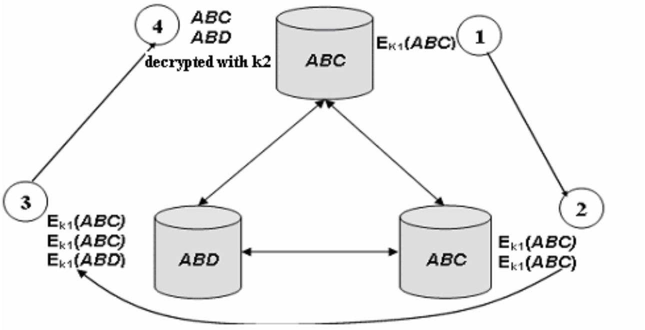


**Fig. 3.** Steps needed for computing the algorithm in [5]

union of a frequent local itemset. We find that this method reduce the number of steps from 6 to 4 to calculate the global candidate item sets as shown in fig. (3) where K1 is private key and k2 public key . In [7] the results showed that this improvement reduces the time of mining process compared to method in [6].

## 3 Proposed New Algorithm

As before we say that the protocol in [6] employs commutative encryption algorithm, so it adds large overhead to the mining process and protocol in [7] improves this by applying a public-key cryptosystem. We can enhance the method in [7] by first rearranged the path to compute the protocol as shown in fig. (4). This can be done by first making two of parties in protocol (one as data mining combiner and one as protocol initiator) and other parties as clients for data mining combiner. This will reduce communication in computing the protocol because firstly communication take time longer than local mining and secondly instead of using two rounds (one round for compute global frequent item sets and one round for compute global support ) we use only one round for computing frequent item sets and global support. Finally improve the time of running data mining algorithm by using Apriori-Tid [10] instead of standard Apriori. This based on tree structure to compute the mining results. Therefore the mining

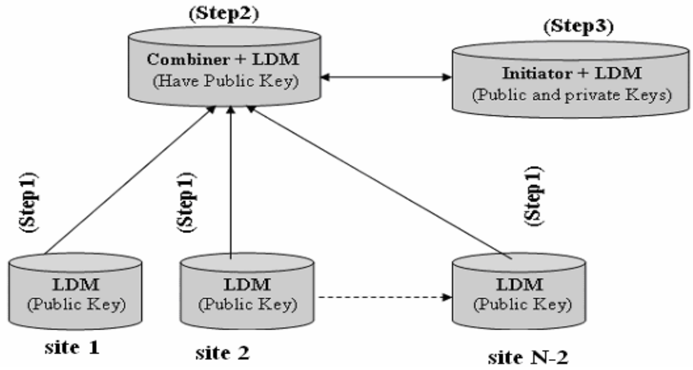


**Fig. 4.** General structure of new proposed algorithm

algorithm takes less time than other because it makes less number of scan over the data to make the mining results.

Step 1: All local data mining (LDM) compute the mining results using fast distributed mining of association rules (FDM) [10] as locally large k-item sets (LLi(k)) and local support for each item set in LLi(k) then Encrypt frequent item sets and support (LLei(k)) then send it to the data mining combiner.
Step 2: The combiner merge all received frequent items and supports with the data mining combiner frequent items and support in encrypted form then send LLe(k) to algorithm initiator to compute the global association rules .
Step3: The algorithm initiator receives the frequent items with support encrypted. The initiator first decrypt it, then merges it with his local data mining result to obtain global mining results L(k), then compute global association rules and distribute it to all protocol parties. The protocol details of our algorithm as in fig. (5).

```
Protocol: Finding large itemsets of size k and global association rules.
Require: $N > 3$ sites one site is algorithm initiator and another is data mining combiner and
other called clients (local data mining) sites numbered $(1..N - 2)$ and we assume negligible
collision probability.
  Step 1: perform local data mining and then encryption of all frequent items sets and local
supports for every frequent item by all clients sites and send it to data mining combiner.
for each site i do
   generate LLi(k) as in FDM algorithm [10]  (step1 and step2)
     $LLei(k) = \varnothing$
  for  each $X \in LLi(k)$ do
           compute  local support for X as Y
           $LLei(k) = LLei(k) \cup \{Ei(X), Ei(Y)\}$
  end for
   send LLei(k) to data mining combiner
end for
  Step2 : data mining combiner merges LLe(k) for every clients with his LLe(k)
generate LLi(k) as in FDM algorithm [10] (step1  and step2)
for each $X \in LLi(k)$ do
          compute local support for X as Y
          $LLei(k) = LLei(k) \{Ei(X), Ei(Y)\}$
end for
merge all LLe(k) of clients and data mining combiner
   send LLe(k) to protocol initiator
  Step 3: Decryption of LLe(k) and compute LLi(k) for the initiator then compute the global
mining results.
for each $X \in LLe(k)$ do
          $LL(k) = D(LLe(k))$
end for
generate LLi(k) as in FDM algorithm [10] (step1 and step2)
for each $X \in LLi(k)$ do
  compute local support for X
end for
eliminates duplicates from the LL(k) and LLi(k)
compute L(k) from LL(k) and LLi(k)
compute association rules with minimum confidence from L(k) and global support of each
item in L(k) and broadcast the results.
```

**Fig. 5.** Proposed new algorithm

Our method reduces the number of steps from four steps to only three steps for any numbers of clients to calculate the global candidate item sets. An example of our protocol shown in fig. (6) where K1 is private key and k2 public key . For the two party cases we can use the protocol without computing global support as above but we use the same method of computing global support as list in [6].

*Proofing* that our algorithm preserves the privacy can be done by using the idea of simulating every thing during the protocol running to know what data every site see in running the protocol [3]. The proof as following:

In step 1: Their is no communication between client sites and results are encrypted and data mining combiner don't have the private key then no privacy loss.

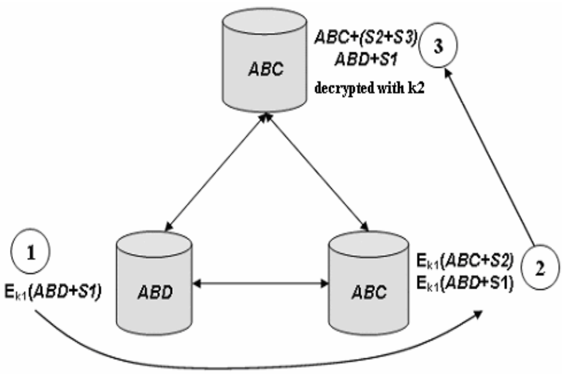


**Fig. 6.** Steps needed for computation of our algorithm

In step 2: Because the results of local data mining are mixed then the initiator can't connect between any data and corresponding site then no privacy loss in this step.

In step 3: finally the initiator compute the final L(k) and publish the final association rules and any site can't deduce from it any information about others. So this protocol can't loss the privacy in the semi honest model and for malicious model the collision found if the initiator and combiner collude with each other and because we choose the initiator and combiner every time running the protocol so we have negligible collision in our protocol.

### 3.1 Performance Metrics

We will consider computation time and communication time for performance metrics and rule quality for accuracy metric.

For measuring the performance of our method we use communication and computation costs as performance metrics. Cost estimation for association rule mining using the method we have presented can be computed as following: The number of sites is N. Let the total number of locally large itemsets be |LLi(k)|, and the number of that can be directly generated by the globally large (k) itemsets be L(k). Let t be the number of bits in the output of the encryption of an itemset. A lower bound on t is

$\log_2(|L(k)|)$; based on current encryption standards t = 512 is a more appropriate value. The total bit-communication cost for the protocol is $O(t^2*|L(k)|* N)$ where L(k) is closer to UiLLei(k) and t is the number of bits in the encryption key. For comparison, The algorithm in [6] need $O(t^3*|[UiLLi(k)|*N^2)$ and algorithm in [7] need $O(t^2*|[UiLLi(k)|*N^2)$ and finally our improved algorithm need $O(t^2*|[UiLLi(k)|*N)$.

For measuring the rule quality. In [12] measuring the quality of data which a privacy preserving algorithm is applied is closely related to the information loss resulting from changing the original data when applying the new algorithm. This measure used to know that if modification affect the quality of the results of mining. In modified algorithm no change made in original data. All we make is encrypting the data in all sites. Then collect it with protocol combiner and initiator. During the protocol we merge the data with each other without affecting it and finally we decrypt this data to make the mining. The original data is returned again with no modification. The mining results now same as if we don't make any change in the distributed mining algorithm[10]. By this proof we say that our algorithm don't loss the accuracy of results and information loss is close to zero.

## 4 Implementation and Results of Proposed Method

We implement our new method and algorithm in [7] using java. Because the distributed association rules mining need the mining run in more than one site, we can use the RMI (remote method invocation) to connect the sites with each other. Our application is two parts one name server and other is client so we have two site works as server. First is the protocol initiator and second is the data mining combiner and we need client for every participant in the protocol. The initiator is responsible of the threshold of the mining algorithm so it need to define the support and confidence and also generate the public key (k2)and private key (k1) used in encryption and decryption in protocol and finally compute the final results. The data mining combiner responsible of combining the results of clients sites and mix the results to make better privacy of user data and every client is responsible of making the local data mining and encrypt the results of mining and send to data mining combiner.

One of the most features of our implantation that our implementation can be extended to any number of sites without modifies the implementation. But in old algorithm to add any new site we need to change the implementation to be extended for this new site. Our test in data that represent based on 0/1 matrix. And using Public-Key Cryptography as in [11]. We use RSA that is useful to fulfill our requirements. By running the new algorithm 75 tests done, and 75 tests for old algorithm. On data bases with different size from 2500 bytes to 2500000 bytes by 15 tests for every data base. The 15 values are very closed to each other. The values listed in table [1] are the average values. Testing done by using P4 (2.8 GHZ) with Java (SDK 1.6). Fig. (7) shows graphically percentage time comparison of our method and old one.

To measure the rule quality we implement the fast distributed mining [10] and make some tests in this algorithm related to our algorithm as in table [2]. We use different data base size 2500, 2500, 50000, 250000 bytes when number of attributes is 50 and 2500000 bytes when number of attributes is 250, confidence = 60 and support = 40.

**Table 1.** Time of new algorithm verse old algorithm

| DB Size | Algorithm in [7] | New Algorithm |
|---|---|---|
| 2500 | 0.110583441 | 0.081683539 |
| 25000 | 1.211517773 | 0.40404254 |
| 50000 | 2.358482587 | 0.722406912 |
| 25000 | 11.6384336 | 3.1813554 |
| 250000 | 109.4533864 | 53.23276793 |

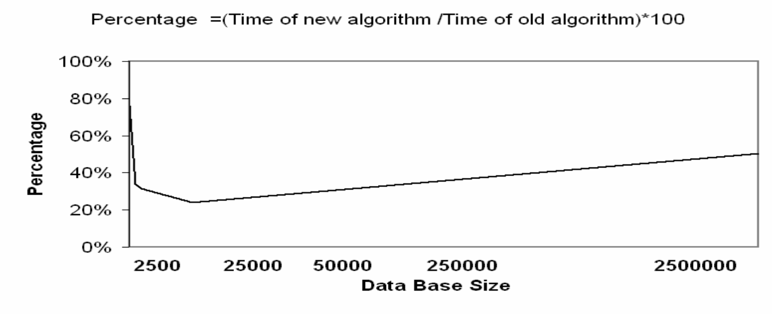


**Fig. 7.** Percentage time of new algorithm related to old algorithm

**Table 2.** Comparison number of rules in new algorithm and old algorithm

| DB Size | Algorithm in [10] | New Algorithm |
|---|---|---|
| 2500 | 161 | 161 |
| 25000 | 23 | 23 |
| 50000 | 20 | 20 |
| 250000 | 18 | 18 |
| 2500000 | 1236 | 1236 |

From the results we can find that new algorithm has a high performance in computations, communications time and accuracy than the algorithm in [7]. This due to the total bit-communication cost for our algorithm is function in (N) site, but the algorithm in [7] is function in ($N^2$) sites. In the same time based on proofing for privacy preserving our algorithm preserves the privacy also. New algorithm is more flexible to extend it to any number of sites without any change in implementation. And also any increase doesn't add more time to algorithm because all client sites perform the mining in the same time so the overhead in communication time only. Also one of the interesting features of this new algorithm is that using Apriori-Tid no need for the database to count the support of any frequent item set after the first pass. New algorithm computes frequent item sets and local support at the same time.

## 5 Conclusion

In this paper we presented privacy preserving association rule mining algorithms of have been recently introduced with the aim of preventing the discovery of sensible information. We modify an algorithm of privacy preserving association rule mining on distributed homogenous data by optimize the communication between sites, modify the mining algorithm and how compute the distributed association rule mining. This modification for the algorithm of computing the distributed association rule mining to preserve the privacy of users. Also an implementation for modified algorithm is presented. From the results obtained we can say that our algorithm is good privacy preserving algorithm and preserve the accuracy with high performance. Our algorithm is more flexible to extend to any number of sites without any change in implementation. And also any increase doesn't add more time to algorithm because all client sites perform the mining in the same time so the overhead in communication time only. The total bit-communication cost for our algorithm is function in (N) sites. Our protocol is base in public key cryptography so we use RSA as example for testing of our algorithm. For future work we can replace it with strong public key cryptography.

## References

1. Ajmani, S., Morris, R., Liskov, B.: A trusted third-party computation service. Technical Report MIT-LCS-TR-847, MIT (May 2001)
2. Agrawal, D., Aggarwal, C.C.: On the design and quantification of privacy preserving data mining algorithms. In: Proceedings of the Twentieth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, Santa Barbara, California, USA, May 21-23, pp. 247–255. ACM, New York (2001)
3. Goldreich, O.: Secure multi-party computation (working draft), `http://www.wisdom.weizmann.ac.il/oded/pp.html`
4. Ioannidis, I., Grama, A.: An efficient protocol for yao's millionaires' problem. In: Hawaii International Conference on System Sciences (HICSS-36), Waikoloa Village, Hawaii, January 6-9 (2003)
5. Lindell, Y., Pinkas, B.: Privacy Preserving Data Mining. Journal of Cryptography, 177–206 (2002)
6. Kantarcioglu, M., Clifton, C.: Privacy-preserving distributed mining of association rules on horizontally partitioned data. IEEE Transactions on Knowledge and Data Engineering Journal 16(9), 1026–1037 (2004)
7. Estivill-Castro, V., Hajyasien, A.: Fast Private Association Rule Mining by a Protocol Securely Sharing Distributed Data. In: Proceedings of the 2007 IEEE Intelligence and Security Informatics (ISI 2007), New Brunswick, New Jersey, USA, May 23-24, pp. 324–330 (2007)
8. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules. In: Proceedings of the 20th International Conference on Very Large Data Bases, Santiago, Chile: VLDB, September 1994, pp. 487–499 (1994)
9. Cheung, D.W.-L., Ng, V., Fu, A.W.-C., Fu, Y.: Efficient mining of association rules in distributed databases. IEEE Transactions on Knowledge and Data Engineering 8(6), 911–922 (1996)
10. Cheung, D.W.-L., Han, J., Ng, V., Fu, A.W.-C., Fu, Y.: A Fast Distributed Algorithm for Mining Association Rules. In: Proc. 1996 Int'l Conf. Parallel and Distributed Information Systems (PDIS 1996), pp. 31–42 (1996)
11. Rivest, R.L., Shamir, A., Adleman, L.: A method for obtaining digital signatures and public-key cryptosystems. Communications of the ACM 21(2), 120–126 (1978)
12. Bertino, E., Fovino, I.N., Provenza, L.P.: A Framework for Evaluating Privacy Preserving Data Mining Algorithms. Data Mining and Knowledge Discovery 11(2), 121–154 (2005)

# SD-Core: Generic Semantic Middleware Components for the Semantic Web*

Ismael Navas-Delgado and José F. Aldana-Montes

Computer Languages and Computing Science Department, University of Málaga,
Málaga, Spain
{ismael,jfam}@lcc.uma.es

**Abstract.** This paper describes a middleware for building Semantic Web applications. The main idea behind this proposal is to help developers build Semantic Web applications by providing them with the main components for this task. This set of components has been implemented and made available through a Web tool (http://khaos.uma.es/SD-Core), and as an installable tool for a Tomcat Web Server. In addition, it has been applied to develop a semantic integration system for biological data sources (http://asp.uma.es/WebMediator).

**Keywords:** Semantic Web, Knowledge-Based Infrastructure.

## 1 Introduction

Semantic Web research has been ongoing since the initial proposal of Tim Berners Lee. Nowadays, this research is producing technology that is being integrated in enterprise applications. In this context, the development of Semantic Web based applications has had to address several problems: choose a component to deal with ontologies, deal with ontology relationships (usually available as ontology alignments), relate non-semantic resources with semantics through annotation tasks, etc. These new issues in software development have caused developers significant problems when estimating the real cost of applications, and reusing existing components.

The essential role of middleware is to manage the complexity and heterogeneity of distributed infrastructures. On the one hand, middleware offers programming abstractions that hide some of the complexities of building a distributed application. On the other, a complex software infrastructure is necessary to implement these abstractions. Instead of the programmer having to deal with every aspect of a distributed application, the middleware takes care of some of them.

Ontologies serve various needs in the Semantic Web, such as storage or exchange of data corresponding to an ontology, ontology-based reasoning or ontology-based navigation. When building a complex Semantic Web application, designers may prefer to combine different existing software modules. Thus, our proposal aims to

* Funded by the ICARO Project, TIN2005-09098-C05-01 (Spanish Ministry of Education and Science), and Applied Systems Biology Project, P07-TIC-02978 (Innovation, Science and Enterprise Ministry of the regional government of the Junta de Andalucía).

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 617–622, 2008.

develop a middleware infrastructure which will use a set of working components to hide details related with semantics from programmers.

An infrastructure is a set of interconnected structural elements that provides the framework to support an entire structure. In the Semantic Web Framework, this is a set of components which provide the basic elements to develop more complex applications.

Although this Framework [1] is in its initial stages (framework design), we have successfully applied our infrastructure to create Semantic Web applications in real scenarios. The Semantic Web Framework has a structure in which applications are described using simple components. Our infrastructure however, describes bigger components because the analysis of the possible Semantic Web applications indicates that some combinations of simple components are shared in all of these applications. The Knowledge Web European project (http://knowledgeweb.semanticweb.org) has designed a Semantic Web Framework, which describes the main elements that Semantic Web applications will require and the type of applications that can be developed by using this framework. Thus, the main aim of this project is to provide components (developed in Universities and Enterprises) that could be useful for developing Semantic Web applications. This approach will assist software developers in finding Semantic Web components and their interconnection.

This framework classifies the components in dimensions, chosen as a result of the developers' experience. The following dimensions are considered: Data and metadata management, Querying and reasoning, Ontology development and management, Ontology customization, Ontology evolution, Ontology alignment, Ontology instance generation and Semantic web services.

Each component is described by defining its dependencies with other components, and then a list of use cases is presented. Each use case describes how several components of the framework can be composed to solve a specific problem. From these use cases and the component dependencies, we can deduce that some blocks of components can be grouped, because they usually act together. In addition, their capabilities are almost the same as the ones provided by our middleware. These common blocks are the Ontology and Data repositories, whose definitions are [1]:

- *Ontology repository component.* This component provides functionalities to locally store and access ontologies and ontology instances.
- *Data repository component.* This component provides functionalities to locally store and access both annotated and un-annotated ontology data.
- *Metadata registry component.* This component provides functionalities to locally store and access metadata information.

These three components are also interesting in that they do not depend on any other component and they can be considered the core of any Semantic Web Application.

In this paper we present the main elements of a middleware for building Semantic Web Applications. The goal of the proposed middleware is to provide useful components for registering and managing ontologies and their relationships, and also to provide metadata about the resources committed or annotated with the ontologies registered, which means a practical step towards building applications in the Semantic Web. The main advantage of using a middleware for the development of Semantic Web applications is that software developers can reuse components, reducing the implementation costs.

In this paper we describe the proposed middleware, showing the characteristics of its generic components (Section 2). Then, we will describe an application use case (Section 2.3). Finally, we will discuss the main advantages of our proposal together with the conclusions reached with respect to the implementation and use.

## 2 Semantic Directory Core

This section presents the generic middleware for the development of Semantic Web applications. The analysis of different architectural proposals and Semantic Web applications makes it clear that a Semantic Web application must have these characteristics: Ontologies are used to introduce semantics; A single, common ontology is not available for most of the domains. Ontology management and alignment is necessary; Resources are annotated with different ontologies, even in the same domain; and Resources need to be located by means of the defined semantics.

Summarizing the list of requirements, we can deduce that ontology and resource managers are necessary components for most of the applications. In addition, we can find rich relationships between ontologies and resources (this is one of the main characteristics of Semantic Web applications), which we can take advantage of when developing Semantic Web applications.

### 2.1 Infrastructure for the Middleware

This infrastructure is based on a resource directory, called Semantic Directory Core, SD-Core (Figure 1). We define the SD-Core as "a set of core elements to build Semantic Web applications, and it is made available as a server to register semantics providing services to query and browse all the registered semantics". In order to formally define the elements that the SD-Core will manage, we have defined the internal elements of the Semantic Directory using metadata ontologies.

Thus, the SD-Core is composed of two inter-related ontologies (OMV [2] and SDMO), which describe the internal semantics of the Semantic Directory (see Figure 1). The main advantage of using metadata ontologies is that this metadata can be managed by tools ranging from a simple OWL parser to a complex ontology reasoner. We use OMV to register additional information about ontologies to help users locate and use them. The metadata scheme contains further elements describing various aspects related to the creation, management and use of an ontology.

SDMO is the ontology in charge of registering information about resources and relationships between these resources and ontologies registered in the Semantic Directory. SDMO and OMV are related by a class included in SDMO, which provides a way of relating resources (SDMO instances) with registered ontologies (OMV instances). The current version of SDMO is composed of five classes: OMV, Resource, Mapping, Similarity and User.

The SD-Core is composed of three interfaces, which tends to be the minimum set of elements for building a wide range of applications for the Semantic Web. These interfaces are made public as Web Services to enable their use in Semantic Web applications.

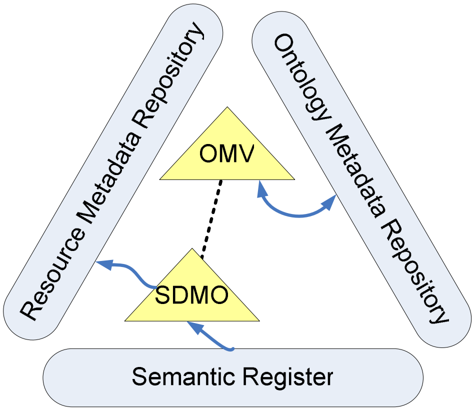


**Fig. 1.** SD-Core interface. The internal elements of SD-Core have been developed as metadata ontologies.

The *Ontology Metadata Repository Interface* is an interface, which offers different types of access to the information related with ontologies registered in the SD-Core. The basic access operation is to search ontologies and instances of the OMV. Where a reasoner is used in the SD-Core the search capabilities will be improved. The following methods are some of those provided to register and browse ontologies: *registerOntology(url,name), getOntology(name), getOntology( url), listOntologies() and listOntologies(concept).*

The *Semantic Register Interface* is in charge of relating resources with several of the registered ontologies. When registering a resource, the interface implementations will generate an instance of the SDMO containing mappings between this resource schema (or ontology) and previously registered ontologies. Thus, it will be possible to take advantage of registered resources using the ontologies registered in the Semantic Directory. Once a resource has been registered the SD-Core monitor (the application in charge of ensuring that all components work correctly) will repetitively test if it is available. In the case where the resource is not available (reachable) it is marked as not available temporarily. Thus, if a user/application asks for resources complying with certain characteristics, un-available resources' URLs will not be returned. If the SD-Core monitor detects that the resource is available, its state is updated.

This interface provides the following methods, which allow the resource owner to register the resource by indicating the mappings manually or by taking advantage of an automatic matching tool: *registerResource(serviceName,url, queryMethod, schemaMethod, capabilityMethod, ontologyName, relationships) and registerResource(serviceName, url, queryMethod, schemaMethod, capabilityMethod)*

The *Resource Metadata Repository Interface* is an interface for registering and accessing information about resources, which provides methods for locating resources based on their URL, name, relationships with domain ontologies, etc. The basic access operations are to search ontologies and instances of the SDMO. The following list briefly describes the main methods provided by this interface:

*getRelatedElements( domainOntology,concept), getSchema(resourceName), getSchema(url), listMappings(resourceName) , listMappings(url), listResources(), listResources(relatedOntology), listResources(relatedOntology, relevantConcepts) and listResources(url).*

### 2.2 Middleware Implementation

SD-Core is accessed through three main components, which tends to be the minimum set of elements for building a wide range of applications for the Semantic Web. These components are responsible for the main task provided by semantic directories and have been developed as three different Java classes, made publicly available as Web Services (developed as an application of Apache Tomcat with Axis). Thus, the installation is as easy as adding these applications to our Web Server. In addition, we have developed an installable version that will include SD_Core in an existing Web Server or in a new one that will be installed if there is no other one available. The overload that SD-Core adds to the Web Server is minimal because the methods developed are lightweight and do not require the installation of any additional applications.

Metadata ontologies are loaded using Jena [3], to establish a memory version of the metadata information. Then, each registration of an ontology or a resource will involve the creation of an instance that is stored in the memory using the Jena Framework.

In order to provide persistence to the system, the metadata ontologies and their instances are stored as a two separate OWL files. These files will be used each time it is necessary to recover the system because of maintenance work or system failures.

### 2.3 Application Example (Semantic Integration)

In order to validate the middleware proposal, we have decided to develop an ontology-based mediator (http://asp.uma.es/WebMediator/), which will take advantage of the SD-Core for dealing with semantics. For this mediator, we propose the use of an ontology which is supposed to formalize a shared and consensus knowledge, the ontology used to integrate the data will be stable. Registering of resources in the Semantic Directory is a key step towards the development of the integration solution, and this task is helped by ontologies. The architecture of the proposed Ontology-Based Mediator is composed of four main components: *Controller, Query Planner,Query Solver and Integrator.*

In this proposal, the sources are made available by publishing them as Web Services (named Data Services). Our primary goal here is to integrate databases accessible via internet pages. In this context, wrappers are an important part of the internal elements of data services. Data services, independently of the development process, are distributed software applications that receive queries in XQuery and return XML documents. In the context of mediator development, the process of registering resources in an SD-Core implies finding a set of mappings between one or several ontologies and the data service schema (usually expressed as an XMLSchema document). These mappings will be the key elements to integrate all the data sources, and these mappings will be the way in which the resource semantics are made explicit.

## 3 Conclusion

This paper describes a middleware for building Semantic Web applications (available at http://khaos.uma.es/SD-Core as a Web Demo version and downloadable installation program). This middleware commits with an architecture that has been described as a set of elements and three interfaces. The implementation of these interfaces will produce different applications, depending on the implementation itself. Thus, we have developed a first implementation and an application based on this infrastructure: a semantic integration application.

The development of this application and its application in a real scenario (protein structure prediction) show how the proposed infrastructure can be used for building real applications.

However, the existence of a list of application types, like the one proposed by the Semantic Web Framework [1], has motivated us to study how to adapt our proposal in other scenarios. Thus, future work to be developed will include the implementation of an SD-Core totally configurable to enable the metadata scheme to be adapted according to the kind of applications and their specific needs.

Finally, the system developed makes use of OWL files to manage metadata which limits its use to small repositories. In order to solve this problem we are developing a new version that uses a persistence system to store the metadata of the SD-Core.

## References

1. Leger, A., Gómez-Pérez, A., Maynard, D., Zyskowski, D., Hartmann, J., Euzenat, J., Dzbor, M., Zaremba, M., del Carmen Suárez-Figueroa, M., García-Castro, R., Palma, R., Dasiopoulou, S., Costache, S., Vitvar, T.: Architecture of the semantic web framework. Technical report (February 2007)
2. Hartmann, J., Sure, Y., Haase, P., Palma, R., del Carmen Suarez-Figueroa, M.: Omv - ontology metadata vocabulary. In: Gil, Y., Motta, E., Benjamins, V.R., Musen, M.A. (eds.) ISWC 2005. LNCS, vol. 3729. Springer, Heidelberg (2005)
3. Carroll, J., Dickinson, I., Dollin, C., Reynolds, D., Seaborne, A., Wilkinson, K.: Jena: Implementing the semantic web recommendations (2003)

# A Knowledge-Based Approach for Answering Fuzzy Queries over Relational Databases

Z.M. Ma and Xiangfu Meng

College of Information Science and Engineering, Northeastern University,
Shenyang, China
mazongmin@ise.neu.edu.cn, marxi@126.com

**Abstract.** Users often have vague or imprecise ideas when searching the database. Thus, they might like to issue fuzzy queries for possibly retrieving. Based on fuzzy sets theory and the knowledge base related to the application domain, this paper proposes an approach translating the fuzzy query into the precise query and extending the query criteria ranges in order to provide approximate answers to the user. The fuzzy condition is first defined by a fuzzy number with membership function, and then is translated into the precise condition by using the α-cut operation of fuzzy number. The tuples satisfying the fuzzy query are finally ranked according to their satisfaction degree.

**Keywords:** Relational database, Fuzzy query, Knowledge base, Ranking.

## 1 Introduction

Database query processing models have always assumed that the user knows what he or she wants and is able to formulate query that accurately expresses the query intentions. However, users often have vague or imprecise ideas when searching the database and thus may like to issue fuzzy queries, which consist of fuzzy terms or fuzzy relations for possibly retrieving. Consider a realtor database for Web applications, for example, which consists of a single table estateDB with attributes (*Price, City, Bedrooms, SchoolDistrict, SqFt, BuildYear* …). Each tuple represents a house for sale in the US. Assume that a potential house buyer searches for houses in this database and would like a new house like that: it is priced close to $350,000 and its surface is between 700 and 800. According to this needs, the fuzzy query can be formulated as follows.

*Q*:- estateDB (Price close to 350000, SqFt between 700 and 800, BuildYear = Recent)

Obviously, this fuzzy query contains a fuzzy term (e.g. "Recent") and a fuzzy relation (e.g. "close to"). However, most databases are still based on the SQL for querying nowadays and cannot handle such a fuzzy query. So it is necessary to translate the fuzzy query into precise query. The user may also be satisfied with the house which surface is slightly lower than 700 or larger than 800. This flexibility extends the query criteria ranges to provide some approximate answers to the user.

Fuzzy values have been employed to model and handle imprecise information in databases since Zadeh introduced the fuzzy sets theory [13]. Tahani [11] firstly advocated the use of fuzzy sets for querying conventional databases, where imprecise

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 623–630, 2008.

conditions inside queries were seen as fuzzy sets. SQLf language proposed by Bosc and Pivert [2], which is a fuzzy extension to SQL, represents a synthesis of the characteristics and functionalities suggested in other previous proposals of flexible query in classical databases, such as Tahani [11], Bosc et al. [3], Wong and Leung [12], and Nakajiam et al. [10]. Also there are some extensions to SQLf [4, 7, 8]. The translation of the fuzzy queries has been presented in [5], which only investigated the fuzzy conditions that contain simple fuzzy terms or fuzzy relation "(not) close to".

Our previous work in [9] has developed the translation of the fuzzy query against relational database, but there is a lack of domain knowledge for constructing membership functions. Furthermore, it is also interesting to rank the query results according to the user's needs and preferences. While some approaches have been proposed for ranking the database query results [1, 4, 6], these approaches mainly focus on the precise query results and most likely hard to rank the fuzzy query results. Extending our previous work in [9], this paper proposes an approach for translating fuzzy queries by leveraging the knowledge base and investigates the ranking method.

The rest of this paper is organized as follows. Section 2 briefly reviews fuzzy sets theory. Section 3 introduces the forms of fuzzy queries. Section 4 presents the knowledge-based approach of fuzzy query translation. Section 5 describes the fuzzy query results ranking approach. The experiment results are presented in Section 6. The paper is concluded in Section 7.

## 2 Fuzzy Sets Theory

Fuzzy data is originally described as fuzzy set by Zadeh [13]. Let $U$ be a universe of discourse. A fuzzy value on $U$ is characterized by a fuzzy set $F$ in $U$. A membership function $\mu_F$: $U \rightarrow [0, 1]$ is defined for the fuzzy set $F$, where $\mu_F(u)$, for each $u \in U$, denotes the membership degree of $u$ in the fuzzy set $F$. A fuzzy number is a fuzzy subset in the universe of discourse $U$ that is both convex and normal.

Let $U$ and $F$ be the same as the above. Then we have the notion of $\alpha$-cut of the fuzzy number. The set of the elements which degrees of membership in $F$ are greater than (greater than or equal to)$\alpha$, where $0 \leq \alpha < 1 (0 < \alpha \leq 1)$, is called the strong (weak) $\alpha$-cut of $F$, respectively, denoted by

$$F_{\alpha+} = \{u \mid u \in U \textit{ and } \mu_F(u) > \alpha\} \text{ and } F_\alpha = \{ u \mid u \in U \textit{ and } \mu_F(u) \geq \alpha\}. \quad (1)$$

## 3 The Forms of Fuzzy Query

The traditional precise query over relational databases is composed of the basic condition $A\theta Y$, where $A$ is an attribute, $\theta$ is the regular operator, and $Y$ is the operand. Similarly, the fuzzy basic condition in a fuzzy query can be described by fuzzy terms or fuzzy relations. Combing fuzzy terms and various regular relations, or combing some fuzzy relations with precise values, the fuzzy conditions are formed.

### 3.1 Fuzzy Terms as Operands

Three kinds of fuzzy terms can be identified: *simple fuzzy term*, *modified fuzzy term*, and *compound fuzzy term* [3]. The fuzzy term can be defined by a fuzzy number with membership function according to the application domain knowledge. Using the fuzzy terms and traditional regular operators, the fuzzy condition with fuzzy operands, which has the form $A\theta\tilde{Y}$, is formed [9]. Here $\tilde{Y}$ is a fuzzy term as the operands.

### 3.2 Fuzzy Relations as Operators

There are three types of fuzzy relations, which are "(*not*) *close to*", "(*not*) *at least*" and "(*not*) *at most*" [9]. Using these fuzzy relations and precise values, the fuzzy condition with fuzzy operators, which has the form $A\tilde{\theta}Y$, is formed. Here, $\tilde{\theta}$ is a fuzzy relation, $Y$ is a precise value, and $\tilde{\theta}Y$ is a fuzzy number with membership function.

### 3.3 Numerical Interval as Operators

For $A\theta Y$, where $A$ is an attribute, $\theta$ is the operator (*not*) *between*, and $Y$ is an interval represented by $[Y_1, Y_2]$, it has the form of "*A* (*not*) *between* $[Y_1, Y_2]$". Then $A\theta Y$ can be executed by DBMS without any translation. But the user's query may also be satisfied by the numerical values nearby the numerical interval. So it is necessary to expand the numerical interval. The expanded interval $[Y_1, Y_2]$ is called as the fuzzy interval. According to [13], the membership function of the fuzzy interval "*between* $[Y_1, Y_2]$" can be defined as follows.

$$\mu_{between[Y_1,Y_2]}(u)=\begin{cases} 0 & , \quad u\le\omega \mid u\ge\delta \\ \dfrac{u-\omega}{Y_1-\omega} & , \quad \omega<u<Y_1 \\ 1 & , \quad Y_1\le u\le Y_2 \\ \dfrac{\delta-u}{\delta-Y_2} & , \quad Y_2<u<\delta \end{cases} \tag{2}$$

The membership function of the fuzzy interval "*between* $[Y_1, Y_2]$" is shown in Fig. 1. Here larger values of $\omega$ correspond to a sharply inclined curve and smaller values of $\omega$ correspond to a flatly inclined curve, while smaller values of $\delta$ correspond to a sharply inclined curve and larger values of $\delta$ correspond to a flatly inclined curve.

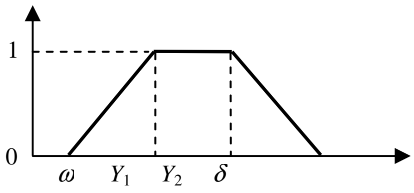


**Fig. 1.** Membership function of the fuzzy interval "*between* $[Y_1, Y_2]$"

## 4 Translation of Fuzzy Query

We first introduce the knowledge base, and then illustrate the fuzzy query translation by using membership functions and α-cut operation of fuzzy numbers.

### 4.1 The Knowledge Base

The knowledge base (KB) stores information needed to translate the fuzzy query into the precise query and relax query criteria. The KB includes the values for constructing membership functions and other information such as the relaxation directions according to the attributes semantics. The KB is maintained by the domain experts. It is composed of the following tables where the primary key of each table is underlined:

**MemFunction** (*FuncName*, *Attribute*, *Table*, *para1*, *para2*, *para3*, *para4*, *Satisf*)
**AttRelaxation** (*Attribute*, *Table*, *FunName*, *Directionrel*, *Lsatisfy*, *Rsatisfy*)

The table **MemFunction** provides the parameter values for constructing membership functions. The attributes *para1*, *para2*, *para3* and *para4* represent the values of corresponding parameters of the membership function, respectively. The attribute *Satisf* represents the threshold of membership function which indicates that the query condition must be satisfied with minimum degree threshold in [0, 1].

The table **AttRelaxation** provides the information for relaxing the query criterion, which includes, for each relaxable attribute and each membership function, the relaxation direction (left and/or right) and the satisfaction type (decreasing or non-decreasing). For example, let the query criterion be “Price close to 350000”. By “relaxation direction is left”, we mean that for the relaxation of this criterion, we only consider values less than 350,000. By “satisfaction type on the left is decreasing”, we mean that the satisfaction degrees for the values being less than 350,000 are less than 1. If we provide the user with a price less than 350,000, he or she is completely satisfied. So the relaxation type is non-decreasing (satisfaction degree is equal to 1).

In order to translate the fuzzy query given in Section 1, we need the tables in the KB as follows (Table 1 and Table 2).

**Table 1.** Instance of table MemFunction

| FunName | Attrbuite | Table | Para1 | Para2 | Para3 | Para4 | Satisf |
|---|---|---|---|---|---|---|---|
| f_Closeto | Price | estateDB | 20000 | - | - | - | 0.8 |
| f_Between | SqFt | estateDB | 600 | 700 | 800 | 900 | 0.8 |
| f_Recent | BuildYear | estateDB | 5 | 10 | 10 | - | 0.8 |

**Table 2.** Instance of table AttRelaxation

| Attribute | Table | FunName | Directionrel | Lsatisfy | Rsatisfy |
|---|---|---|---|---|---|
| Price | estateDB | f_Closeto | left, right | non-decr | decr |
| SqFt | estateDB | f_Between | left, right | decr | non-decr |
| BuildYear | estateDB | f_Recent | right | non-decr | decr |

### 4.2 Translation of Fuzzy Query

The idea of translating fuzzy query is to convert a fuzzy basic condition into a precise condition. The truth that a threshold should be chosen for the fuzzy query makes it possible to do that.

**Translation of fuzzy query with fuzzy operands.** For fuzzy condition "$A\,\theta\tilde{Y}$ *WITH* $\alpha$", let α be a given threshold, which is considered as a retrieval threshold and α∈ [0, 1]. Generally speaking, the smaller the value of α is, the more the number of tuples satisfying the given condition and being provided by the system is.

It is clear that the α-cut of $\tilde{Y}$ is an interval. Let $Y_\alpha$= [*a*, *b*], then "$A=\tilde{Y}$ *WITH* $\alpha$" can be translated into "*A* ≥*a AND A*≤*b*". Consider the fuzzy condition "*BuildYear = Recent WITH 0.8*". According to [9] and the domain knowledge in KB, the fuzzy term "*Recent*" on the universe of discourse can be defined by

$$\mu(u)=\begin{cases} 1 & , \quad u\le 5 \\ [1+(\dfrac{u-5}{2})^2]^{-1} & , \quad 5<u<10 \\ 0 & , \quad u\ge 10 \end{cases} \tag{3}$$

The result of 0.8-cut operation on fuzzy term "*Recent*" is [1, 6]. Then the fuzzy condition can be translated into "*BuildYear*≥*1 AND BuildYear* ≤*6*".

**Translation of fuzzy query with fuzzy operators.** For fuzzy condition "$A\,\tilde{\theta}Y$ *WITH* $\alpha$", let $\alpha$ be a given threshold and the α-cut of $\tilde{\theta}Y$ is with the form [*a*, *b*] (i.e. $\tilde{\theta}Y_\alpha$=[*a*, *b*]). Then "$A\,\tilde{\theta}Y$ *WITH* $\alpha$" can be translated into "*A* ≥*a AND A*≤*b*".

Consider the fuzzy condition "*Price close to 350000 WITH 0.8*". According to [5] and the domain knowledge in KB, the fuzzy number "*close to 350000*" on the universe of discourse can be defined by

$$\mu_{close\ to\ 350000}(u)=\frac{1}{1+\left(\dfrac{u-350000}{20000}\right)^2} \tag{4}$$

The result of 0.8-cut operation on "*close to 350000*" is [340000, 360000]. Then the fuzzy condition can be translated into "*Price*≥*340000 AND Price*≤*360000*".

**Extending numerical interval.** For the fuzzy condition "*A between* [$Y_1$, $Y_2$] *WITH* $\alpha$", let α be a given threshold and $[Y_1, Y_2]_\alpha = [y_1, y_2]$ (note that $y_1<Y_1$ and $y_2>Y_2$). Then "*A between* [$Y_1$, $Y_2$] *WITH* $\alpha$" can be translated into "$A \ge y_1$ AND $A\le y_2$".

Consider the fuzzy condition "*SqFt between 700 and 800 WITH 0.8*". According to Equation (2) and the domain knowledge in KB, the fuzzy interval "*between 700 and 800*" can be defined by

$$\mu_{between[700,800]}(u)=\begin{cases} 0 & , \quad u\le 600 \mid u\ge 900 \\ \dfrac{u-600}{100} & , \quad 600<u<700 \\ 1 & , \quad 700\le u\le 800 \\ \dfrac{900-u}{100} & , \quad 800<u<900 \end{cases} \tag{5}$$

The result of 0.8-cut operation on fuzzy interval "*between* [*700,800*]" is [680,820]. Then the fuzzy condition can be translated into "*SqFt*≥*680 AND SqFt*≤*820*".

As discussed above, the fuzzy conditions are translated into precise conditions and the query criteria are extended consequently. As a result, the approximate answer items can be retrieved. In general, a fuzzy query may result in too many relevant answer items and it is necessary to rank the fuzzy query results.

# 5 Fuzzy Query Results Ranking

In this section we first measure the attribute weight by leveraging the workload, and then propose the membership degree ranking approach of fuzzy query results.

## 5.1 Attribute Weight Assignment

Database *workloads – log of past user queries*, have been shown as being a good source for implicitly estimating the user interest. The workload information can help determine the frequency with which database attributes are often specified by users and thus may be important to new users [1]. Therefore, we determine the attribute weight by evaluating the frequency of occurrence of the attribute specified in queries in the workload. In another words, the more frequent the attribute is, the more important the attribute is for most users and thus the higher the weight is. Let $F(A_i)$ be the frequency of occurrence of the attribute $A_i$ specified in queries in the workload. Let $N$ be the number of queries in the workloads. Then we have the Equation (6) for assigning the attribute weight.

$$W(A_i) = (F(A_i)+1)/N \tag{6}$$

## 5.2 Membership Degree Ranking Approach

Let $Q$ be a fuzzy query over the relational database $R$ and $T$ be an answer tuple for $Q$. Also let the set of attributes $X = \{X_1, \ldots, X_k\} \subseteq A$ be the set of attributes specified by the fuzzy conditions $<C_1,\ldots, C_k>$ in $Q$. Then, the membership degree of the answer tuple $T$ to the fuzzy query $Q$ can be defined as

$$D(T, Q)=\sum_{i=1}^{k} W(X_i)\times\mu_{C_i}(t_i) \tag{7}$$

Here $k$ is the number of fuzzy conditions in $Q$, $X_i$ is the attribute specified by the fuzzy condition $C_i$ in $Q$, $W(X_i)$ is the attribute weight of attribute $X_i$, and $\mu_{C_i}(t_i)$ is the membership degree of the value $t_i$ of attribute $X_i$ to the fuzzy condition $C_i$.

The main idea of MDR (membership degree ranking) approach is that the query results are ranked according to their membership degrees to the fuzzy query. The larger the membership degree is, the higher the ranking score is for the result tuple.

# 6 Experiments

The experiments aim at testing the ranking quality of MDR algorithm for fuzzy query results. We set up a realtor database *estateDB* (*Price, City, Bedrooms, Bathrooms,* *Location, SqFt, BuildYear, View*) containing 25,000 tuples extracted from Yahoo! RealEstate. For building a workload, we request twenty people to provide us with queries that they would execute if they want to buy a house. We collect 200 queries for this database and these queries are used as the workload.

Besides MDR described above, we implement two other ranking methods, which are RANDOM and PIR, to compare with MDR. In the RANDOM ranking model, the tuples in the query results are presented to the user in a random order. In PIR [4], given a result tuple *t*, its ranking score is composed of a global score and a conditional score. This approach addresses the similar problem (i.e., ranking of many query results) as MDR does. Our approach differs from that in [4] as follows: MDR considers the value difference of the specified attributes and it is also supported by the knowledge base when ranking, while PIR only focuses on the unspecified attributes.

For formally comparing the ranking precision of the various ranking functions, we retain 3080 tuples and generate 11 test queries. For each query $Q_i$, we generate a set $H_i$ of 30 tuples likely to contain a good mix of relevant and irrelevant tuples to the query. Finally, we present the queries along with their corresponding $H_i$'s to each user in our study. Each user's responsibility is to rank the top 10 tuples as the relevant tuples that they prefer most from the 30 unique tuples collected for each query. We use a standard collaborative filtering metric *R* proposed in [1] to measure ranking quality. Fig. 2 shows the ranking precision of different ranking algorithms.

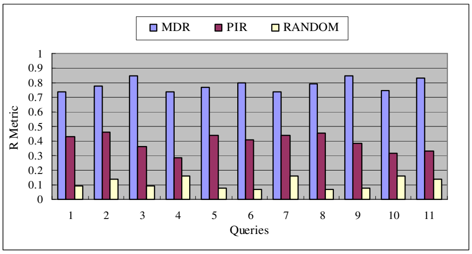


**Fig. 2.** The ranking precision of different ranking algorithms for each test query

It can be seen that MDR greatly outperforms both PIR and RANDOM. The averaged ranking precision of MDR is 0.78 while PIR is 0.39.

## 7 Conclusions

In this paper, we have proposed an approach for translating fuzzy queries over relational databases. Based on fuzzy set theory, this approach uses membership functions, knowledge base and α-cut operation of fuzzy numbers to assist the translation process. The tuples satisfying the fuzzy query are ranked according to their satisfaction degree. The experiment results have shown that the translation of fuzzy query and ranking method can effectively meet user's needs and preferences. How to make the knowledge base running more efficient without experts' assistants is our further work.

**Acknowledgments.** Work is supported by the Program for New Century Excellent Talents in University (NCET-05-0288) and in part by the MOE Funds for Doctoral Programs (20050145024).

## References

1. Agrawal, S., Chaudhuri, S., Das, G., Gionis, A.: Automated ranking of database query results. ACM Transactions on Database Systems 28(2), 140–174 (2003)
2. Bosc, P., Pivert, O.: SQLf: a relational database language for fuzzy querying. IEEE Transactions on Fuzzy Systems 3(1), 1–17 (1995)
3. Bosc, P., Galibourg, M., Hamon, G.: Fuzzy querying with SQL: extensions and implementation aspects. Fuzzy Sets Systems 28, 333–349 (1988)
4. Bordogna, G., Psaila, G.: Extending SQL with customizable soft selection conditions. In: 20th ACM Symposium on Applied Computing, pp. 1107–1111. ACM Press, Santa Fe (2005)
5. Chen, S.M., Jong, W.T.: Fuzzy query translation for relational database systems. IEEE Transactions Systems, Man, and Cybernetics-Part B: Cybernetics 27(4), 714–721 (1997)
6. Chaudhuri, S., Das, G., Hristidis, V.: Probabilistic information retrieval approach for ranking of database query results. ACM Transaction on Database Systems 31(3), 1134–1168 (2006)
7. Goncalves, M., Tineo, L.: SQLf flexible querying extension by means of the norm SQL2. In: 10th IEEE International Conference on Fuzzy Systems, pp. 473–476. IEEE Press, New York (2001)
8. Goncalves, M., Tineo, L.: SQLf3: an extension of SQLf with SQL3 features. In: 10th IEEE International Conference on Fuzzy Systems, pp. 477–480. IEEE Press, New York (2001)
9. Ma, Z.M., Yan, L.: Generalization of strategies for fuzzy query translation in classical relational databases. Information and Software Technology 49(2), 172–180 (2007)
10. Nakajima, H., Sogoh, T., Arao, M.: Fuzzy database language and library: Fuzzy extension to SQL. In: 2nd IEEE International Conference on Fuzzy Systems, pp. 477–482. IEEE Press, San Francisco (1993)
11. Tahani, V.: A conceptual framework for fuzzy querying processing: a step toward very intelligent databases systems. Information Processing Management 13, 289–303 (1997)
12. Wong, M., Leung, K.: A fuzzy database-query language. Information Systems 15(5), 583–590 (1990)
13. Zadeh, L.A.: Fuzzy sets. Information and Control 8(3), 338–353 (1965)

# An Algorithm for Discovering Ontology Mappings in P2P Systems

Giuseppe Pirrò[1], Massimo Ruffolo[2,3], and Domenico Talia[1,3]

[1] University of Calabria, [2] ICAR-CNR, [3] Exeura s.r.l., Rende, Italy
{gpirro,talia}@deis.unical.it, ruffolo@icar.cnr.it

**Abstract.** Ontology Mapping is mandatory for enabling semantic interoperability among different agents and services making use of different ontologies. The ontology mapping problem becomes more critical in P2P systems since: (i) the number of different ontologies can dramatically increase; (ii) ontology mapping must be performed on the fly and only on parts of ontologies contextual to a specific interaction in which the peers are involved; (iii) complex mapping strategies (e.g., structural mapping) cannot be exploited since peers are not aware of one another's ontologies. Hence, specific techniques have to be designed. This paper presents and evaluate the SEmantiC COordinator (*SECCO*) ontology mapping algorithm that addresses the abovementioned issues by adopting a mapping strategy based on the evaluation of three different similarity measures: syntactic, lexical and contextual.

**Keywords:** Peer-to-Peer, ontology mapping, semantic mapping, semantic web.

## 1 Introduction

The Semantic Web aims at providing Web resources (e.g., web pages, documents) with supplementary meaningful information (i.e., metadata) for improving and facilitating their retrieval and enabling their automatic processing by machines. Ontologies are key enablers towards this "new" Web of semantically rich resources. Ontologies can be exploited to give "shared conceptualizations" of knowledge domains and make "explicit" and machine-understandable the meaning of the terminology adopted [10]. Most ontology languages used today are based on XML (e.g., RDF(S) [12], OWL [15]) thus making ontologies exploitable in different class of systems such as Peer-to-Peer applications [20]. From a recent interview to Tim Berners-Lee [1] emerges that semantic-based data sharing is expected to begin in controlled environments smaller than the World Wide Web as for instance: enterprise networks and small-medium Peer-to-Peer (P2P) networks.

In distributed environments, is not feasible to have a single (and universally accepted) ontology describing a knowledge domain, but rather there will be different ontologies created w.r.t "the point of view" of their designers. In order to promote (semantic) interoperability between these different perspectives about the world, it is necessary to ensure "reciprocal understanding". This problem has been a core issue of recent ontology research and in literature is referred to as the Ontology Mapping

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 631–641, 2008.

Problem (OMP). OMP concerns discovering correspondences (*aka* mappings) between entities belonging to different ontologies (i.e., a *source* and a *target* ontology). The OMP becomes more challenging in P2P networks for the following reasons: (i) the number of possible overlapping ontologies dramatically increases since each peer could have its own ontology that reflects peer's needs and interests; (ii) ontology mapping must be performed "on the fly" and only on the parts contextual to the specific interaction in which peers are involved; (iii) peers are unaware of one another's ontologies and therefore the amount of ontological information exploitable to discover mappings is quite limited. In the literature, several approaches to OMP are available. However, these approaches, generally designed to work offline, cannot completely address the specific requirements for P2P networks above mentioned. A recent survey on ontology mapping [5] refers only one work on mapping systems designed for P2P networks. That points out how the importance of the OMP in P2P environment is still underexplored.

This paper presents the novel ontology mapping algorithm named SEmantiC COordinator (*SECCO*) designed for facing the OMP in P2P systems. Given a *seeker* and a *provider* peer, *SECCO* computes *concept mappings* among concepts contained in peer ontologies. It adopts a new mapping strategy based on the exploitation of three similarity measures evaluated by means of three different matchers: syntactic, lexical and contextual. The mapping strategy of *SECCO* is based on the idea of concept *context* exploited in the *contextual matcher*. Given an ontology concept, the context is constituted by its properties and the set of other concepts with which it is directly related. This way a context is an agile encoding of the amount of structural information needed to evaluate similarities among ontology concepts. By adopting the notion of context, *concept mappings* are obtained just evaluating how a seeker concept fits in the context of provider concepts. This new mapping strategy, which complies with the contextual theory of meaning [14], avoids performing complex structural analysis of peer ontologies. Moreover, it allows building *semantic links* among peers that can be exploited in several classes of semantic P2P applications (e.g., semantic search, semantic query routing). The most important feature of *SECCO*, mainly due to the *contextual matcher*, is the ability to face the OMP from a perspective that enables accuracy and fastness in P2P networks. Experimental results show that *SECCO* provides satisfactory results (in terms of quality of mappings).

The remainder of this paper is organized as follows. Section 2 introduces the terminology adopted in the rest of the paper and describes the *SECCO* P2P ontology mapping algorithm. Section 3 evaluates the system on different ontologies. Section 4 reviews related work. Section 5 draws some conclusions, sketches future work and concludes the paper.

## 2 The *SECCO* Ontology Mapping Algorithm

### 2.1 Preliminary Definitions

**Definition 1 (Ontology).** An ontology is basically composed of two parts: schema and instances. For the scopes of this paper, we only consider the ontology schema defined as a five-tuple of the form: $O=\langle C,\leq_C,R,\leq_R,\varphi_R\rangle$ consisting of a set of concepts

$C$ and a set of relations $R$ respectively arranged in hierarchies by means of the partial orders $\leq_C$, $\leq_R$. The signature $\varphi_R$: $R \to C \times C$ associates each relation $r \in R$ with its domain $dom(r)=\pi_1(\varphi_R(r))$ and range $range(r)= \pi_2(\varphi_R(r))$.

Fig. 1 shows two excerpts of (online available) ontologies. The first ontology *(Ka)* describes research projects while the second *(Portal)* the content of a Web portal.

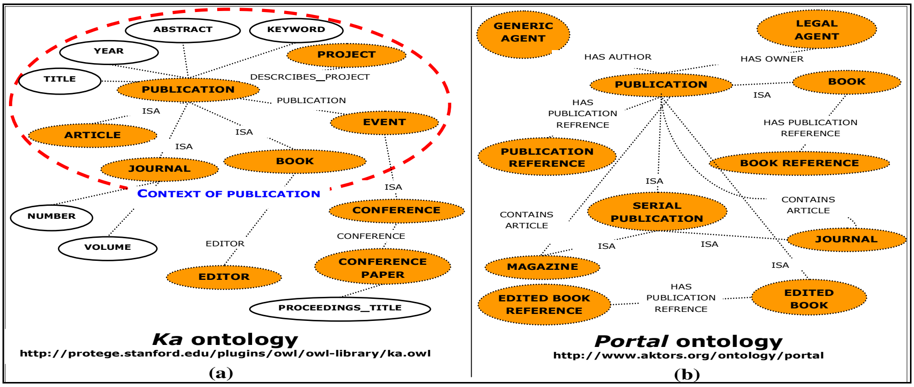


**Fig. 1.** Excerpts of the *Portal* and *Ka* ontologies defining the concept *Publication*. Concepts are represented as filled oval while properties as empty ovals.

**Definition 2 (Seeker and Provider peer).** A *seeker* peer is a semantic peer that sends a *request* over the P2P network to *provider* peers that receive the *request*, execute *SECCO* and return *concept mappings*.

**Definition 3 (Concept Context).** Let $O$ be an ontology, the set $ctx(c)=C_{range}\cup C_{dom}\cup R_{dp}$ is the *context* of the *concept* $c\in C$ where: (i) $C_{range}$ and $C_{dom}$ are the sets of concepts for which given an object property $r\in R$ respectively hold the following conditions $c\in dom(r)\wedge c_{range}\in range(r)$ and $c_{dom}\in dom(r)\wedge c\in range(r)$; (ii) $R_{dp}$ is the set of names of data type properties for which the following condition holds $c\in dom(r)\wedge d\in range(r)$ where $r\in R$ is a data type property and d is a data type [15]. This way for each concept, a context contains its properties and other concepts directly related to it. For the excerpt of ontology in Fig. 1 (a) the context of the *Publication* concept contains the following elements *ctx(Publication)={TITLE, YEAR, ABSTRACT, KEYWORD, PROJECT, EVENT, BOOK, JOURNAL, ARTICLE}*.

**Definition 4 (Request).** Let $O$ be an ontology, a *request* is a two-tuple of the form $RQ=\langle c, ctx(c)\rangle$ where $c\in C$ is a concept belonging to a *seeker* peer ontology and $ctx(c)$ is the *context* of $c$. For the excerpt of ontology in Fig. 1 (a) the request takes the following form: $RQ=\langle Publication, ctx(Publication)\rangle$.

**Definition 5 (Concept Mapping).** Let the *seeker* peer request $RQ=\langle s, ctx(s)\rangle$ and the *provider* peer ontology $O_p$, a *concept mapping* $M$ between each provider concept $p\in C_p$ and the *seeker* peer concept $s\in C_s$ is a set of 3-tuples of the form $\langle s, p, \sigma\rangle$ where $\sigma\in [0,1]$ is the similarity value between $s$ and $p$.

**Definition 6 (Similarity Measure).** Let a *seeker* and a *provider* ontology $O_s$ and $O_p$, a request $RQ_s=\langle s,ctx(s)\rangle$ and the set $CTX_p$ composed by two-tuples of the form $\langle p_j,ctx(p_j)\rangle$ where $\forall p_j\in C_p$ the set $ctx(p_j)$ is the context of $p_j$; the similarity between the couples of concepts $s\in C_s$ and $p\in C_p$ is computed by the following function:

$$\mathrm{sim}(s,\ p) : f\,(\mathrm{sim}_{\mathrm{syn}}\,(s, p),\ \mathrm{sim}_{\mathrm{lex}}\,(s, p),\ \mathrm{sim}_{\mathrm{con}}\,(s, p_p)\,)\rightarrow [0.1]$$

where $sim_{syn}(s,p)$: $C_s\times C_p \rightarrow [0,1]$ is the *syntactic similarity*, $sim_{lex}(s,p)$: $C_s\times C_p \rightarrow [0,1]$ is the *lexical similarity*, $sim_{con}(s,p)$: $RQ_s\times CTX_p \rightarrow [0,1]$ is the *contextual similarity* and $f$ is a function exploited to combine results (e.g., weighted sum).

## 2.2 The SECCO Algorithm

We consider a P2P network in which each peer owns an *ontology* exploited to conceptualize its view on a particular knowledge domain. Each peer in the network plays a twofold role: (i) *seeker* peer, when it sends a request to the network; (ii) *provider* peer, when it executes locally the *SECCO* algorithm. Provider peers use the *SECCO* algorithm to obtain *concept mappings* between a *seeker* concept *s* and their ontology concepts. Whenever a *provider* peer receives a request, it runs *SECCO* with an input

```
                          The SECCO algorithm
Input: An input I=<c_s, ctx(c_s), O, T_h, w_s, w_l, w_c> where O=⟨C,R⟩
Output: The concept mapping M
Method:
  M=∅;
  for each c∈C do
   sim_syn=evaluate_syntactic_similarity(c_s,c); //see Section 2.3.1
   sim_lex=evaluate_lexical_similarity(c_s,c); //see Section 2.3.2
   ctx(c)=extract_context(c,O); // see below
   sim_con=evaluate_contextual_similarity(c_s,Ctx(c_s),c,Ctx(c)); //see Fig. 4
   sim=(w_s*sim_syn+w_l*sim_lex+w_c*sim_con); //overall similarity value
   if sim>T_h then
      m.s=c_s
      m.p=c;
      m.σ=sim;
      M=M∪m;
   end-if
  end-for
  return M;
                       Function extract_context
Input: An ontology O=⟨C,R⟩ and a concept c∈C
Output: The context ctx(c)
Method:
      ctx_c=∅;
      for each c_c∈C do
          for each r_c∈R do
              if ∃r_c(c,c_c)|∃r_c(c_c,c) then
                 if(datatype(r_c))
                     ctx_c= ctx_c∪{r_c} //where r_c is the name of the property
                 end-if
                 else
               ctx_c=ctx_c∪{c_c}
              end-if
          end-for
      end-for
      return ctx(c)=⟨ctxc⟩;
```

**Fig. 2.** The *SECCO* algorithm in pseudo-code

of the following form: $I=<s, ctx(s), O_p, T_h, w_s, w_l, w_c>$ where: *s* is the *seeker* peer concept; *ctx(s)* is the context of *s*; $O_p$ is the *provider* peer ontology; $T_h \in [0,1]$ is a threshold value for similarity between couple of concepts, if the similarity is above the threshold the couple of concepts appear in the mapping; $w_s$, $w_l$, $w_c$ are used to weight the contribution of the *syntactic*, *lexical* and *contextual* similarities. The pseudo-code of the *SECCO* algorithm is shown in Fig. 2.

Fig. 3 depicts the overall approach. A *seeker* peer issues an information request by picking a concept along with the related context from its ontology. This request reaches *provider* peers that run the *SECCO* algorithm. The *Combiner* module of *SECCO* exploits the similarity function (see Definition 6) to combine similarity values provided by the matchers and filters the results according to the threshold $T_h$. On termination, *SECCO* returns a *concept mapping* (see Definition 5) to the *seeker* peer that will stored it in the *mapping store*. Mappings are exploitable in several classes of semantic P2P applications such as semantic search, semantic-based query routing, and community formation.

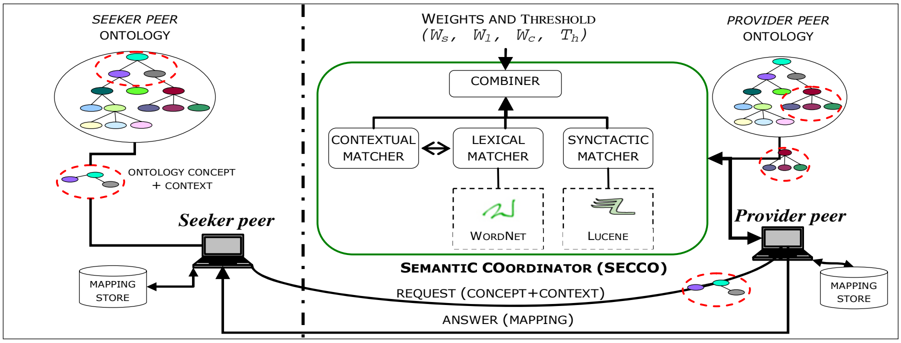


**Fig. 3.** The *SECCO* architecture and usage scenario

## 2.3 Individual Matchers: The Building Blocks of SECCO

A single heuristic evaluation of the similarity degree between couples of ontology concepts cannot exploit all the types of information (e.g., lexical and structural) encoded in ontology entities. For instance, a linguistic matcher (e.g., string matching) can compare the names of ontology entities but cannot exploit structural information expressed by the relations that entities have with their neighbors. Therefore, as done in other systems (e.g., [7]), *SECCO* adopts a mapping strategy that exploits multiple similarity measures computed by different matchers which have been explicitly considered for their suitability in the P2P context.

### 2.3.1 The Syntactic Matcher

The syntactic matcher implements the function *evaluate_syntactic_similarity* and relies on the Lucene Ontology Matcher (LOM) based on the Lucene search engine library (http://lucene.apache.org). LOM [17] exploits different sources of linguistic information (e.g., local name, comments, and labels) present in ontology entities. In particular, each entity belonging to a *source* ontology is transformed into a *virtual* *document* by exploiting the concept of Lucene Document. Virtual documents are stored into an index stored in main memory. Ontology mappings are derived by using values of *target* ontology entities as search arguments against the index created from the source ontology. Similarity values are computed by exploiting the scoring schema implemented in Lucene. Further details along with complete experimental results can be found in [17].

### 2.3.2 The Lexical Matcher

The *lexical matcher*, which implements the function *evaluate_lexical_similarity*, allows interpreting the semantic meaning of entities to be compared by exploiting the knowledge contained in WordNet [13]. The aim of the *lexical matcher* is to assess relatedness between ontology entities. Relatedness is a special case of semantic similarity. In facts, while semantic similarity only considers the relations of Hyponymy/Hypernymy among WordNet *synsets*, semantic relatedness, takes into account a broader range of semantic relations. To consider both aspects, we combine a measure of semantic similarity with a measure of relatedness. We evaluated a variety of similarity and relatedness measures on a set of similarity ratings provided by humans collected by an online experiment. More details are available at the GridLab web site: (http://grid.deis.unical.it/similarity). We adopt the following formula that is an adaptation of the Jiang and Conrath metric (J&C) [11] for computing semantic similarity between two concepts $c_s$ and $c_p$:

$$sim(c_s,c_p) = 1 - \frac{IC(c_s) + IC(c_p) - 2 * IC(sub(c_s,c_p))}{2} \quad (1)$$

where values of Information Content (IC) [18] are obtained by the WordNet structure as described in [19] and $sub(c_s, c_p)$ indicates the concept that subsumes $c_s$ and $c_p$.

As relatedness metric, we exploit the gloss vector metric [16] whose rationale is to compare *synsets'* glosses for assessing relatedness. In particular, this approach exploits "second order" vectors for glosses, that is, rather than just matching words that occur in glosses, the words in the gloss are replaced with co-occurrence (extracted from a corpus) vectors. Therefore, each gloss is represented by the average of its word vectors. Hence, pairwise comparisons can be made between vectors to measure relatedness between the concepts they represent. If $v_1$ and $v_2$ are the gloss vectors for $c_s$ and $c_p$, their relatedness in computed as follows:

$$relat\ (c_s,c_p) = \frac{v_1 * v_2}{\left|\overrightarrow{v_1}\right| * \left|\overrightarrow{v_2}\right|} \quad (2)$$

Overall, the lexical similarity is computed as follows:

$$lex(c_s,c_p) = w_s * sim(c_s,c_p) + w_r * relat(c_s,c_p) \quad (3)$$

where $w_s$ and $w_r$ are the contributions of semantic similarity and relatedness to the final lexical similarity value. From experimental evaluation, we obtained optimal results by equally weighting the two contributions.

### 2.3.3 The Contextual Matcher

The aim of the *contextual matcher* is to implement the *evaluate_contextual_similarity* function (see Fig. 2) exploited to refine similarity values assessed by the syntactic and/or lexical matcher. It advances a contextual approach to semantic similarity that builds upon Miller at al. [14] definition in terms of the interchangeability of words in contexts. Contexts help to refine the search of correct mappings since they intrinsically contain both information about the domains in which concepts to be compared are used and their structure in terms of properties and neighbors concepts. Contexts represent possible patterns of usage of concepts and the contextual matcher is based on the idea that similar concepts have similar patterns of usage. The underlying idea can be summarized as follows: a concept $c_s$ in a context $ctx(c_s)$ not similar to a concept $c_p$ in a context $ctx(c_p)$ will likely fit bad into $ctx(c_p)$ as well as $c_p$ will do in $ctx(c_s)$. If the two concepts can be interchangeably used, that is, fit well in each other's contexts, therefore they can be considered similar. In order to quantify how well a concept fits in a context, we calculate the lexical similarity among the concept and all the concepts in the considered context and take the average value. Overall, the contextual similarity is computed by exploiting the following similarity indicators: (i) *s2s*, which indicates how the seeker concept fits in the seeker context; *s2t*, which indicates how the seeker concept fits in the provider context; *t2t*, which indicates how the provider concept fits in the provider context; *t2s*, which indicates how the provider concept fits in the seeker context. These indicators are obtained by exploiting the *evaluate_how_it_fits function* and combined to obtain an overall similarity value (see Fig. 4). It is worth noting that this strategy aims at taking into account structural information about concepts on a local basis, that is, by only considering properties and nearest neighbors concepts in the taxonomy. This is justified by the following reasons: (i) complex matching strategies (e.g., graph matching strategies [21]) considering the entire ontologies to be matched cannot be applied since a provider peer is not aware

```
Function evaluate_contextual_similarity
Input: Two concepts $c_1$ and $c_2$ and their contexts ctx($c_1$) and ctx($c_2$)
Output: A numerical value $sim_{con} \in [0,1]$ representing the contextual similarity between the concepts $c_1$ and $c_2$
Method:
    s2s = evaluate_how_it_fits($c_1$, ctx($c_1$));
    s2t = evaluate_how_it_fits($c_1$, ctx($c_2$));
    t2s = evaluate_how_it_fits($c_2$, ctx($c_1$));
    t2t = evaluate_how_it_fits($c_2$, ctx($c_2$));
    $sim_{con}$ = ((1-||s2s-t2t|-|s2t+t2s||)); //overall similarity value
    return $sim_{con}$
Function evaluate_how_it_fits
Input: A concepts c and a context ctx(x)= $\langle C_x \rangle$
Output: A numerical value $m \in [0,1]$ representing the fitness between the concept c and the context ctx(x)
Method:
    for each $c_e \in C_x$ do
        T = sum(evaluate_lexical_similarity(c,$c_e$));
    end-for
    return m=T/|ctx(x)|;
```

**Fig. 4.** The pseudo-code of the *evaluate_contextual_similarity* function

of the whole ontology of the *seeker* peer; (ii) a complete mapping among peers ontologies is not required; they only need to map their part of ontologies contextual to the interaction in which they are involved expressed through a request.

# 3 Experimental Evaluation

This section discusses results of experiments we carried out. In Section 3.1 we compare *SECCO* with H-Match [3,4] on the two excerpts of ontologies depicted in Fig. 1. Whereas Section 3.2 discusses the evaluation of *SECCO* on the OAEI tests (http://oaei.ontologymatching.org/2006/).

## 3.1 Experiment 1*:* Comparing *SECCO* with H-Match

We assume that the ontology depicted in Fig. 1 (a) belongs to a *seeker* peer while that in Fig. 1 (b) to a *provider* peer. Moreover, we consider *Publication* as *source* concept. The input *I* of *SECCO* is: *I=< Publication, ctx(Publication), Portal, 0 ,0.1, 0.6,0.3>*. We do not set a threshold value, since we want to find *1:n* mappings. Moreover, we put more emphasis on the semantic features of *SECCO* since the two ontologies do not have labels and comments; therefore the amount of information exploitable by *LOM* is limited. Results obtained by *SECCO* and H-Match are reported in Table 1.

**Table 1.** Comparison between *SECCO* and H-Match

| **Ka concept** | **Portal concept** | *SECCO* | | | | **H-Match [3,4]** | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | *Matcher* | | | | | | | |
| | | Syn | Lex | Ctx | **Total** | *Surface* | *Shallow* | *Deep* | *Intensive* |
| *Publication* | *Publication* | 1 | 1 | 0.697 | **0.909** | 1 | 0.738 | 0.804 | 0.781 |
| *Publication* | *Book* | 0 | 0.823 | 0.199 | **0.553** | 0.8 | 0.618 | 0.66 | 0.639 |
| *Publication* | *Journal* | 0 | 0.767 | 0.221 | **0.526** | 0.64 | 0.522 | 0.553 | 0.538 |
| *Publication* | *Magazine* | 0 | 0.737 | 0.088 | **0.468** | 0.8 | 0.618 | 0.649 | 0.634 |
| *Publication* | *Edited Book* | 0 | 0.401 | 0.674 | **0.443** | 0.64 | 0.522 | 0.564 | 0.543 |
| *Publication* | *Publication Reference* | 0.3 | 0.549 | 0.118 | **0.395** | 0.64 | 0.553 | 0.574 | 0.550 |
| *Publication* | *Book Reference* | 0 | 0.549 | 0.118 | **0.365** | 0.64 | 0.553 | 0.573 | 0.549 |
| *Publication* | *Edited Book Reference* | 0 | 0.549 | 0.118 | **0.365** | 0.64 | 0.553 | 0.563 | 0.542 |

These examples show the suitability of the *lexical matcher*, that allows discovering mapping in a *semantic* way. In fact, by considering the analyzed couples of concepts only from a syntactic point of view we would obtain similarity values equal to 0 apart from the couples *Publication (Ka)* and *Publication* (*Portal*) and *Publication (Ka)* and *Publication Reference* (*Portal*). It is worth noting that *SECCO* and H-Match obtained the higher similarity value on the couple *Publication* (*Ka*) and *Publication* (*Portal*).

## 3.2 Experiment 2*:* Evaluating *SECCO* on the OAEI Benchmark Set of Tests

This section describes the evaluation of *SECCO* on four real-life ontologies contained in the OAEI 2006 test suite. In this evaluation, we want discover 1:1 mappings.

Moreover, we set the threshold value (i.e., $T_h$) to 0.51. To evaluate SECCO, we compare it with H-Match and with other algorithms not designed to tackle the OMP in P2P networks. For each of these ontologies the OAEI organizers provide a reference alignment. We computed measures of *Precision* (*P*), *Recall* (*R*) and *F-measure* (*F-m*) [6] between the reference alignment and results obtained by *SECCO*. Notice that *SECCO*, even being designed for P2P networks, can be exploited to compare entire ontologies by reiterating the process described in Section 2 for each concept in the *source* ontology (i.e., the reference ontology referred to as 101 in the OAEI tests).The results of this evaluation are shown in Table 2.

**Table 2.** Average results obtained by some ontology mapping algorithms on the OAEI 2006 real-life ontologies, as reported in [9]

| | *SECCO* | *Jhu/apl* | *Automs* | *Falcon* | *RiMOM* | *H-Match* |
|---|---|---|---|---|---|---|
| **Precision** | 0.81 | 0.18 | 0.91 | 0.89 | 0.83 | 0.78 |
| **Recall** | 0.81 | 0.50 | 0.70 | 0.78 | 0.82 | 0.57 |
| **F-Measure** | 0.81 | 0.26 | 0.79 | 0.83 | 0.82 | 0.65 |

*SECCO* obtained an average *P* of 0.81, an average *R* of 0.81 and an average *F-m* of 0.81. Among the compared algorithms, *SECCO* is one of the most precise. It is slightly outmatched by Automs and Falcon. In addition, results in terms of *R* and *F-m* are satisfactory. However, we want to underline the following aspect. *SECCO* cannot exploit structural aspects of ontologies in the whole (for the reasons explained in the Introduction). It encompasses structural information of concepts to be mapped only on a local basis through the notion of context. Conversely, most of the presented approaches have a solid (and complex) structural matching strategy (for instance Falcon relies on the GMO-matcher [21]). Despite these limitations, *SECCO* obtains very good results meaning that our strategy based on contexts is reasonable. The comparison with H-Match shows how *SECCO* is better in terms of *P*, *R* and *F-m*.

## 4 Related Work

To date, several ontology mapping algorithms have been proposed (see [5] for a survey). Actually, only a few systems address the ontology mapping problem in open environments. Here we shortly discuss some approaches that share features with *SECCO*. The CtxMatch algorithm [2] aims at discovering mappings between Hierarchical Categories (HCs). It relies on WordNet for interpreting the correct sense of concepts on the basis of the context in which they appear. Therefore, it performs a transformation of the concepts to be compared in Description Logics axioms that are exploited to reduce the problem of discovery mappings to a SAT problem. CtxMatch, similarly to *SECCO,* implements a semantic based approach since it relies on WordNet. However, the main difference between these systems is that CtxMatch focuses on matching HCs. H-Match [3,4] is an algorithm for dynamically matching concepts in distributed ontologies. H-Match allows for different kinds of matching depending on the level of accuracy needed. The system aims at supporting knowledge sharing and ontology-addressable content retrieval in peer-based systems, thus it is the system nearest to *SECCO.* However, there exist several differences between these systems:

(i) the *lexical matcher* of H-Match is based on an ad-hoc thesaurus whereas *SECCO* exploits WordNet and in particular all the kind of semantic relations among synsets; (ii) the *contextual matcher* of *SECCO* is based on a new strategy which complies with the contextual theory of meaning defined by Miller et al. [14]; (iii) the *syntactic matcher* of *SECCO* exploits several sources of linguistic information (e.g., names, labels) of ontology entities. In the literature, there are also some Semantic P2P applications sharing common characteristics with *SECCO*. SWAP (Semantic Web and Peer to Peer) [8] aims at combining ontologies and P2P for knowledge management purposes. In SWAP, mappings between peer ontologies are dynamically obtained by exploiting techniques based on lexical features, structure and instances of ontologies. However, a comparison between the two approaches is not possible since SWAP has not been evaluated on public benchmarks such as the OAEI tests.

## 5 Concluding Remarks and Future Work

This paper described *SECCO*, an algorithm for ontology mapping in P2P systems. The main problem we faced is that in a P2P scenario we cannot adopt sophisticated structural matching strategies that require knowing the whole ontologies to be compared. Therefore, we designed an ad-hoc mapping strategy. We adopted the notion of *context*, defined as a concept along with its properties and nearest neighbors concepts. Through contexts, we aim at encoding the amount of structural information needed in a particular request. We compare contextual information of different concepts by the "how it fits" strategy that is basically founded on the idea that two concepts are similar if they fit well in each other's context. This strategy is supported by a *lexical matcher* and, in order to exploit all the linguistic information of ontology entities, a *syntactic matcher*. In the paper, we proved through experimental results the suitability of the algorithm. As future work, we aim at exploiting mappings discovered by *SECCO* for semantic search, semantic-based query routing, and community formation.

## References

1. Berners-Lee, T.: The Semantic Web: An interview with Tim Berners-Lee. Consortium Standards Bulletin 4(6) (June 2005)
2. Bouquet, P., Serafini, L., Zanobini, S.: Semantic coordination: a new approach and an application. In: Fensel, D., Sycara, K.P., Mylopoulos, J. (eds.) ISWC 2003. LNCS, vol. 2870, pp. 20–23. Springer, Heidelberg (2003)
3. Castano, S., Ferrara, A., Montanelli, S., Racca, G.: From Surface to Intensive Matching of Semantic web Ontologies. In: Proc. of WEBS 2004, Zaragoza, Spain, pp. 140–144 (2004)
4. Castano, S., Ferrara, A., Montanelli, S.: H-MATCH: an Algorithm for Dynamically Matching Ontologies in Peer-based Systems. In: Proc. of SWDB, Berlin, Germany, pp. 231–250 (2003)
5. Choi, N., Song, I., Han, H.: A survey on Ontology Mapping. SIGMOD Record 35(3), 34–41 (2006)
6. Do, H., Melnik, S., Rahm, E.: Comparison of schema matching evaluations. In: Proc. GI-Workshop Web and Databases, Erfurt, Germany, pp. 221–237 (2002)

7. Ehrig, M., Sure, Y.: Ontology Mapping – an integrated approach. In: Bussler, C.J., Davies, J., Fensel, D., Studer, R. (eds.) ESWS 2004. LNCS, vol. 3053, pp. 76–91. Springer, Heidelberg (2004)
8. Ehrig, M., Tempich, C., Broekstra, J., Van Harmelen, F., Sabou, M., Siebes, R., Staab, S., Stuckenschmidt, H.: SWAP: Ontology-based Knowledge Management with Peer-to-Peer Technology. In: Proc. of WOW, Luzerne, Switzerland (2003)
9. Euzenat, J., Mochol, M., Shvaiko, P., Stuckenschmidt, H., Šváb, O., Svátek, V., van Hage, W.R., Yatskevich, M.: Results of the Ontology Alignment Evaluation Initiative 2006. In: Proc. of OM 2006, Athens, Georgia, USA (2006)
10. Gruber, T.R.: A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition 5(2), 199–220 (1993)
11. Jiang, J., Conrath, D.: Semantic similarity based on corpus statistics and lexical taxonomy. In: Proc. of ROCLING X, Taiwan (1997)
12. Klyne, G., Caroll, J.J.: Resource Description Framework (RDF): Concepts and abstract Syntax. W3C Recommendation (February 10, 2004), http://www.w3.org/TR/rdf-concepts/
13. Miller, G.: WordNet An On-line Lexical Database. International Journal of Lexicography 3(4), 235–312 (1990)
14. Miller, G.A., Charles, W.G.: Contextual Correlates of Semantic Similarity. Language and Cognitive Processes 6(1), 1–28 (1991)
15. Patel-Schneider, P.F., Hayes, P., Horrocks, I.: OWL Web Ontology Language Semantic and Abstract Syntax, http://www.w3.org/TR/owl-semantics/
16. Patwardhan, S., Pedersen, T.: Using WordNet-based context vectors to estimate the semantic relatedness of concepts. In: Proc. of EACL 2006 workshop, pp. 1–8 (2006)
17. Pirrò, G., Talia, D.: An approach to Ontology Mapping based on the Lucene search engine library. In: Proc. of SWAE 2007, Regensburg, Germany, pp. 407–412 (2007)
18. Resnik, P.: Using information content to evaluate semantic similarity in a taxonomy. In: Proc. of IJCAI 1995, Montréal, Québec, Canada, pp. 448–453 (1995)
19. Seco, N., Veale, T., Hayes, J.: An intrinsic information content metric for semantic similarity in WordNet. In: Proc. of ECAI, Valencia, Spain, pp. 1089–1090 (2004)
20. Staab, S., Stuckenschmidt, H.: Semantic Web and Peer-to-Peer. Decentralized Management and Exchange of Knowledge and Information. Springer, Heidelberg (2006)
21. Wei, H., Ningsheng, J., Yuzhong, Q., Yanbing, W.: GMO: A Graph Matching for Ontologies. In: Proc. of K-Cap 2005, Banff, Canada, pp. 43–50 (2005)

# Privacy Risks in Trajectory Data Publishing: Reconstructing Private Trajectories from Continuous Properties[⋆]

Emre Kaplan, Thomas B. Pedersen, Erkay Savaş, and Yücel Saygın

Faculty of Engineering & Natural Sciences
Sabanci University, Istanbul, Turkey

**Abstract.** Location and time information about individuals can be captured through GPS devices, GSM phones, RFID tag readers, and by other similar means. Such data can be pre-processed to obtain trajectories which are sequences of spatio-temporal data points belonging to a moving object. Recently, advanced data mining techniques have been developed for extracting patterns from moving object trajectories to enable applications such as city traffic planning, identification of evacuation routes, trend detection, and many more. However, when special care is not taken, trajectories of individuals may also pose serious privacy risks even after they are de-identified or mapped into other forms. In this paper, we show that an unknown private trajectory can be reconstructed from knowledge of its properties released for data mining, which at first glance may not seem to pose any privacy threats. In particular, we propose a technique to demonstrate how private trajectories can be re-constructed from knowledge of their distances to a bounded set of known trajectories. Experiments performed on real data sets show that the number of known samples is surprisingly smaller than the actual theoretical bounds.

**Keywords:** Privacy, Spatio-temporal data, trajectories, data mining.

## 1 Introduction

Information about our location is being collected via an ever-increasing number of devices and by an increasing number of parties, e.g. private companies and public organizations. Phone companies can track our movements via our cellphones. Banks register time and location information for our financial transactions we performed using our credit cards. A growing number of RFID tags are being used to give us access to, e.g., parking spaces or public transportation. Considering the current trend, there is no doubt that the amount of spatio-temporal data being collected will increase drastically in the future. From the point of view of data-analysis, the availability of all this information gives us the

[⋆] This work was partially funded by the Information Society Technologies Programme of the European Commission, Future and Emerging Technologies under IST-014915 GeoPKDD project.

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 642–649, 2008.

ability to find new and interesting patterns about how people move in the public space. For instance, such patterns will be useful in solving the growing traffic problems in many metropolitan areas. On the other hand, collection of all these time and location pairs of individuals enables anyone, who observes the data, to reconstruct the movements (the trajectory) of others with a very high precision. There is a growing concern about this serious threat to privacy of individuals whose whereabouts are easily monitored and tracked. Legal and technical aspects of such threats were highlighted at a recent workshop on mobility, data mining, and privacy [3].

In this paper we consider the following scenario: A malicious person wishes to reconstruct the movements (the "target trajectory") of a specific individual. The malicious person does not know the trajectory itself, but only various properties of the trajectory, such as the average speed, a few points visited, or the average distance between the target trajectory and a few trajectories known to the malicious person. We propose a concrete algorithm which can reconstruct the target trajectory from this information.

Despite privacy concerns, many techniques were proposed to mine useful patterns from trajectories. Some of the very recent results are [5,7,8,10] where in [5] the authors mine for temporal patterns of the form $a \rightarrow^t b$ meaning that $t$ is the typical time to travel from location $a$ to location $b$. Their algorithm needs to know what points of interests the trajectories pass through, and at which time intervals. In [7] the authors give a clustering algorithm which considers sub-trajectories. The main observation is that sub-parts of trajectories may follow interesting common patterns, while the trajectories as a whole may be very different from each other. In [8] authors give a method for finding "hot-routes" in a given road network, which can help us in traffic management.

In all the algorithms mentioned above different properties of the trajectories are needed. Some methods only need the mutual distances between trajectories, some need the exact trajectories, and others only need to know at what times the trajectories pass through certain areas of interest. In this paper, we show how, even very little, information is enough to recover the movement behavior of an individual. In particular we demonstrate how an unknown trajectory can be almost entirely reconstructed from its distance to a few fixed trajectories.

Previous work on spatio-temporal data privacy include anonymization in location based services. Some of the recent work include [9,2]. However, they do not deal with trajectory data. Techniques for trajectory anonymization were recently proposed in [1] but privacy risks after data release were not considered. In another recent work, privacy risks due to distance preserving data transformations were identified [13], however spatio-temporal data was not addressed.

Contributions of this work can be summarized as follows: 1) We demonstrate that trajectories can be reconstructed very precisely with very limited information using relatively simple methods. In particular we show that for a real world dataset of bus trajectories in Athens, we can reconstruct an unknown trajectory with 1096 sample points by knowing its distance to only 40-50 known trajectories. This is in sharp contrast to the 2193 known distances which would be needed to solve the corresponding system of equations to find the unknown trajectory. 2) We propose a method which can reconstruct trajectories from a very wide range of continuous properties (cf. Section 2); the method of known distances is only a special case. Our method is optimal in the sense that it will eventually find a candidate which exhausts all the information available about the unknown trajectory.

## 2 Trajectories and Continuous Properties

In their most general form trajectories are paths in space–time. In practice, however, trajectories are collected with GPS devices, or other discrete sampling methods. A discrete trajectory is a polyline represented as a list of sample-points: $T = ((x_1, y_1, t_1), \ldots, (x_n, y_n, t_n))$. We write $T_i$ to represent the $i$th sample-point $(x_i, y_i, t_i)$. In most of this paper we think of a trajectory as a column-vector in a large vector-space. We use calligraphic letters to refer to the vector representation of a trajectory. The vector representation of a trajectory $T$ is: $\mathscr{T} = (x_1, y_1, t_1, \ldots, x_n, y_n, t_n)^T \in \mathbb{R}^{3n}$. In this case $\mathscr{T}_i$ is the $i$th element of the vector (i.e. $\mathscr{T}_1 = x_1, \mathscr{T}_2 = y_1, \ldots, \mathscr{T}_{3n} = t_n$).

In this paper we assume that trajectories are 1) are *aligned*[1] and 2) have constant sampling rate ($t_{i+1} - t_i = c$, for some constant $c$). Algorithms for ensuring these conditions can be found in [6]. In consequence we discard the time component and represent a trajectory as a list of $(x, y)$ coordinates (or a vector in $\mathbb{R}^{2n}$).

A trajectory $\mathscr{T}$ can posses many properties which are of interest in different situations, such as maximum and average speed of a trajectory, closest distance to certain locations, duration of longest "stop", or percentage of time that $\mathscr{T}$ moves "on road". In this work we show how any property of $\mathscr{T}$ which can be expressed as a continuously differentiable function $f : \mathbb{R}^{2n} \to \mathbb{R}$ can be used to reconstruct $\mathscr{T}$. All the examples given above are continuously differentiable properties of $\mathscr{T}$.

The experiments in Section 5 are performed by using an important property of trajectories, namely the distance from an unknown trajectory $\mathscr{T}$ to a fixed trajectory, $\mathscr{T}'$. When using a continuously differentiable norm to compute the distance between $\mathscr{T}$ and $\mathscr{T}'$ we obtain a continuously differentiable property of $\mathscr{T}$; e.g. $\Delta_{\mathscr{T}'}(\mathscr{T}) = d(\mathscr{T}', \mathscr{T})$ is continuously differentiable. Several distance measures for trajectories have been proposed [11], but in the experiments in this paper we focus on Euclidean distance:

$$\|\mathscr{T} - \mathscr{T}'\|_2 = \sqrt{\sum_{i=1}^{2n} |\mathscr{T}_i - \mathscr{T}_i'|^2}, \tag{1}$$

[1] Two trajectories are aligned if they have the same sampling times and the same number of sample points.

## 3 Reconstructing Trajectories

In this paper we consider how a malicious person can find an unknown trajectory, $X$, with as little information as possible. Any information we have about $X$ may improve our ability to reconstruct $X$; a car does not drive in the ocean, and rarely travels at a speed of more than 200 km/h. With a sufficient number of known properties of $X$, the trajectory can be fully reconstructed. If, for example, $2n$ linear properties of $X$ are known, we have a system of $2n$ linear equations. Solving these $2n$ equations gives us the exact unknown trajectory. The number of linear properties we need to know, however, is at least as large as the number of coordinates in the trajectory itself. If only $m \ll 2n+1$ linear properties are known, the solution will be in a $(2n-m)$-dimensional subspace, at best. When the candidate can only be restricted to a subspace, it can be arbitrarily far away from $X$. If the known properties are non-linear, finding a solution to the corresponding equations, even if sufficient number of properties is known, may even become infeasible.

As seen from this discussion, a method which can approximate the unknown trajectory with considerably fewer known properties than coordinates is needed. The method presented in the next section is an important step in this direction.

In the rest of this paper we limit our study to information about Euclidean distance between the unknown trajectory and $m \ll 2n+1$ known trajectories, and leave it to future work to include other properties of trajectories. The method we propose in the next section, however, can easily be extended to handle any continuously differentiable property. Thus, the problem addressed in the rest of this paper is as follows: Given $m$ trajectories, $\mathscr{T}_1, \ldots, \mathscr{T}_m$, and $m$ corresponding positive real values $\delta_i, \varepsilon_i$, where

$$\delta_i = \|\mathscr{X} - \mathscr{T}_i\| + e_i, \tag{2}$$

for unknown error-terms $e_i$, $|e_i| \leq \varepsilon_i$, and unknown trajectory $\mathscr{X}$, our task is to find an approximation $\mathscr{X}'$ which minimizes the distance $\|\mathscr{X} - \mathscr{X}'\|$.

A natural measure of success of a reconstruction method is the distance $\|\mathscr{X} - \mathscr{X}'\|$. However, this distance depends on the coordinate system of the dataset, and thus tells us very little about the efficiency of the reconstruction method itself. Notice that a naïve approach to estimating $\mathscr{X}$ would be to set $\mathscr{X}'$ to the trajectory $\mathscr{T}_i$ with the smallest distance $\delta_i$. Any meaningful method should give a solution which is closer to $\mathscr{X}$ than $\delta_i$. Thus, we define the success-rate as

$$SR(\mathscr{X}') = 1 - \frac{\|\mathscr{X} - \mathscr{X}'\|}{\delta_{min}}, \tag{3}$$

where $\delta_{min} = \min_i(\delta_i)$ is the smallest given distance. The success-rate is 1 if the method finds $\mathscr{X}$ precisely, 0 if it returns the closest known trajectory, and finally negative if what it does is worse than just returning the closest known trajectory.

To find the unknown trajectory, we need a method which gives meaningful results, even when insufficient amount of information is given. However, the best we can hope for, is to find a candidate trajectory which has the same properties as the properties we know about $\mathscr{X}$. If, for instance, the only information we have about $\mathscr{X}$ is that it is a car driving at an average speed of 50 km/h in Athens, then any $\mathscr{X}'$ which moves along the roads of Athens at 50 km/h is a possible solution. We thus want to minimize the difference between the given properties of $\mathscr{X}$, and the corresponding properties of the candidate $\mathscr{X}'$; in our case, the distances to the known trajectories. To this end, we define the "error" of a candidate $\mathscr{X}'$ as

$$E(\mathscr{X}') = \sum_{i=1}^{n} \left(\|\mathscr{X}' - \mathscr{T}_i\| - \delta_i\right)^2. \tag{4}$$

A natural way to solve this problem is to see it as an optimization problem, which is the essence of our method described in detail in the next section.

## 4 Our Method

We adopt steepest descent (gradient descent search) algorithm to find a candidate with minimum error.

The error-function (4) has value 0 exactly when the candidate trajectory is at distance $\delta_i$ to the known trajectory $\mathscr{T}_i$, for all $i \in \{1, \dots, n\}$. Furthermore, since (4) is a positive valued function, the target trajectory is a global minimum. There may, however, be more than one global minimum, as well as several local minima; but any zero of the error-function exhausts the knowledge we can possibly have about the unknown trajectory. Recall that gradient descent algorithm finds a zero of a positive and continuously differentiable function $E$ as follows

1. Choose a random point, $x_0$, in the domain of $E$.
2. Iteratively define $x_{i+1} = x_i - \gamma \nabla E(x_i)$, for some step-size $\gamma > 0$.
3. When $x_{i+1} = x_i$ ($\nabla E(x_i) = 0$) a (local) minimum has been reached. If $E(x_i) = 0$ we have a global minimum (since $E$ is non-negative), and we stop. Otherwise, we restart at step 1.

The reader may notice that the success-rate as defined in Section 3, with an upper bound of 1, can be an arbitrary negative number and a lower bound for the success-rate may be hard to compute. With the gradient descent method, however, we give a lower bound on the success-rate in Theorem 1.

**Theorem 1.** *Any trajectory $\mathscr{X}'$ with $E(\mathscr{X}') = 0$ has success-rate*

$$SR(\mathscr{X}') \geq 1 - \frac{2\delta_{max} + \varepsilon_{max}}{\delta_{min}}, \tag{5}$$

*where $\delta_{max} = \max_i(\delta_i)$ is the largest given distance, and $\varepsilon_{max}$ is the corresponding error bound.*

*Proof.* By the sub-additivity of the Euclidean norm, $\|\mathscr{X}' - \mathscr{X}\| \leq \|\mathscr{X}' - \mathscr{T}_i\| + \|\mathscr{T}_i - \mathscr{X}\| \leq \delta_i + (\delta_i + \varepsilon_i)$, for all $i \in \{1, \dots, n\}$. Let $\delta_{max} = \max_i(\delta_i)$ be the largest given distance, and $\varepsilon_{max}$ be the corresponding error bound, then $\|\mathscr{X}' - \mathscr{X}\| \leq 2\delta_{max} + \varepsilon_{max}$, and thus $SR(\mathscr{X}') \geq 1 - (2\delta_{max} + \varepsilon_{max})/\delta_{min}$. □

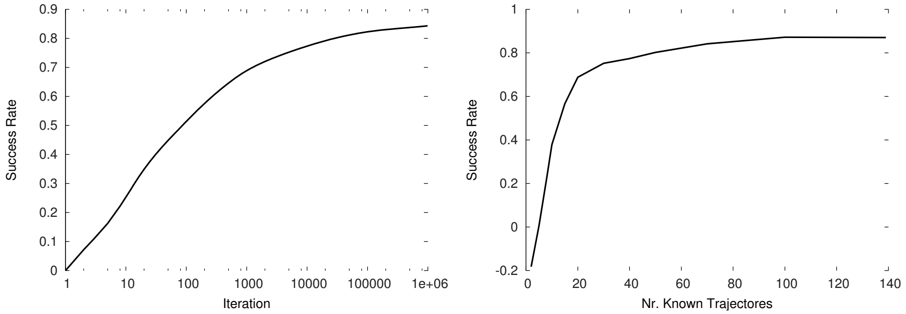
# 5 Experimental Results

Our reconstruction method has been tested on a dataset of GPS data from school busses in Athens[4,12]. The dataset contains 145 trajectories each with 1096 $(x, y)$ sample points. The trajectories are recorded with samples approximately every half minute on 108 different days. For the purpose of our tests we assume that the trajectories are perfectly aligned. In all tests throughout this section, the only property used is Euclidean distance between the target trajectory and some known trajectories. No other property is known to the malicious person.

For the purpose of testing the reconstruction method described in Section 4 we implemented a limited version. In the implementation the step-size $\gamma$ is set to one, and the implementation does not restart if a local maxima, or saddle point is reached. Even though time is not a primary concern in this work, we remark that it takes approximately 8 minutes to run the reconstruction method with 40 known trajectories for 50.000 iterations on a 1.7 GHz laptop on the dataset described below.

Figure 1(a) shows the convergence speed of our reconstruction method. The success-rate is an average value obtained from 15 runs of the test with 50 known trajectories, where the target trajectory is selected at random in each of the 15 runs. The $x$-axis shows the number of iterations in log-scale. Note that in these experiments our reconstruction method finds a candidate which is close to the best it can ever find after approximately 50.000 iterations.

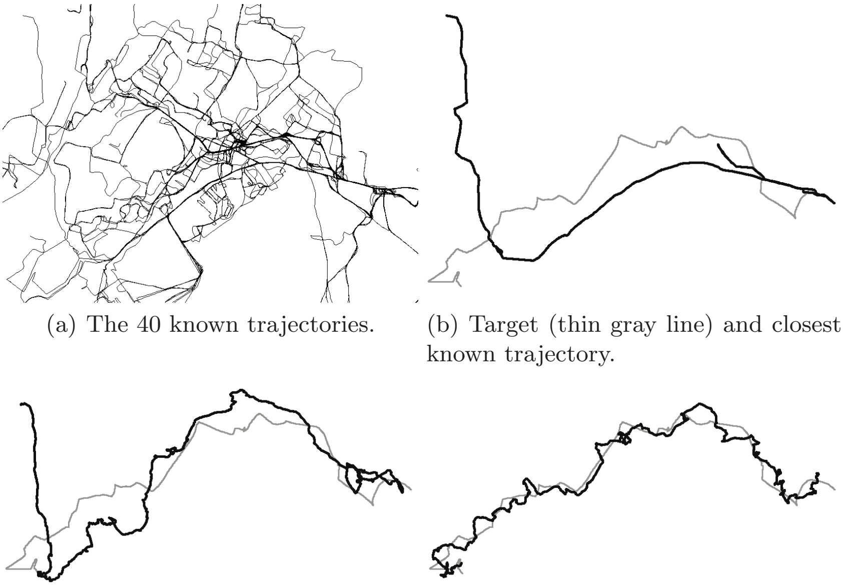
Figure 1(b) shows the success-rate attainable for different numbers of known trajectories. Each sample is the average success-rate of 60 tests with 40 known trajectories, each running for 50.000 iterations. Both target and known trajectories are chosen at random in each test. The graph shows that with less than 5 known trajectories, our reconstruction method is "destructive" (the success-rate is negative); but with 8 known trajectories the success-rate grows already to 0.23. After 100 known trajectories, the success-rate stops growing.



(a) Success-rate vs. number of iterations. The x-axis is in log-scale (Average of 15 experiments with 50 known trajectories).

(b) Success-rate vs. number of known trajectories (Each sample is the average of 60 experiments run for 50.000 iterations).

**Fig. 1.** Success-rate



(a) The 40 known trajectories.

(b) Target (thin gray line) and closest known trajectory.

(c) Success-rate 0.50

(d) Success-rate 0.90

**Fig. 2.** Evolution of the candidate trajectory

Figure 2 shows the evolution of a candidate in one experiment. A candidate with a success-rate of 0.9 clearly shows the whereabouts of the target. However, it must be noted that a success-rate of 0.9 may give a different visual impression for other datasets. We note that for the Athens dataset, most of the trajectories have large overlapping segments (main streets of Athens).

## 6 Conclusion and Future Work

In this paper we present a method for finding an unknown trajectory from knowledge of continuous properties of the trajectory. Our method is optimal in the sense that it will eventually find a candidate which exhausts all the information available about the unknown trajectory.

Our experiments show that unknown private trajectories with 1096 sample points can be reconstructed with an expected success-rate of 0.8 by knowing the distance to only 50 known trajectories. Reconstructing the trajectory perfectly with "tri-lateration" would require 2193 known trajectories.

Adding other known properties such as average speed may improve our method. Knowing the topology of the landscape in which the trajectory is lying is also likely to improve the results of our method, since many false positives will have altitudes which indicate that the candidate "moves through hills". As future work, we will investigate the effects of such properties. We assumed that noise is limited to a known interval. A more realistic model of noise is to let the noise be chosen according to a Gaussian distribution. The present model can handle this to a certain extent using the 99.9% confidence interval as the known limited interval. However, preliminary experiments along these lines suggest that it is better to redesign the "interval function" to handle Gaussian noise.

## References

1. Abul, O., Bonchi, F.: Never walk alone: Uncertainty for anonymity in moving objects databases. In: The 24th International Conference on Data Engineering (ICDE 2008) (2008)
2. Bettini, C., Mascetti, S., Wang, X.S., Jajodia, S.: Anonymity in location-based services: Towards a general framework. In: MDM, pp. 69–76 (2007)
3. First interdisciplinary workshop on mobility, data mining and privacy, rome, italy (February 2008), http://wiki.kdubiq.org/mobileDMprivacyWorkshop/
4. Frentzos, E., Gratsias, K., Pelekis, N., Theodoridis, Y.: Nearest neighbor search on moving object trajectories. In: Bauzer Medeiros, C., Egenhofer, M.J., Bertino, E. (eds.) SSTD 2005. LNCS, vol. 3633, pp. 328–345. Springer, Heidelberg (2005)
5. Giannotti, F., Nanni, M., Pinelli, F., Pedreschi, D.: Trajectory pattern mining. In: KDD 2007: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 330–339. ACM, New York (2007)
6. Gusfield, D.: Efficient methods for multiple sequence alignment with guaranteed error bounds. Bulletin of Mathematical Biology 55(1), 141–154 (1993)
7. Lee, J., Han, J., Whang, K.: Trajectory clustering: a partition-and-group framework. In: SIGMOD 2007: Proceedings of the 2007 ACM SIGMOD international conference on Management of data, pp. 593–604. ACM, New York (2007)
8. Li, X., Han, J., Lee, J.-G., Gonzalez, H.: Traffic density-based discovery of hot routes in road networks. In: Papadias, D., Zhang, D., Kollios, G. (eds.) SSTD 2007. LNCS, vol. 4605, pp. 441–459. Springer, Heidelberg (2007)
9. Mokbel, M.F., Chow, C.-Y., Aref, W.G.: The new casper: A privacy-aware location-based database server. In: ICDE, pp. 1499–1500 (2007)
10. Nanni, M., Pedreschi, D.: Time-focused clustering of trajectories of moving objects. Journal of Intelligent Information Systems 27(3), 267–289 (2006)
11. Needham, C.J., Boyle, R.D.: Performance evaluation metrics and statistics for positional tracker evaluation. In: Third International Conference on Computer Vision Systems, ICVS 2003, pp. 278–289 (2003)
12. http://www.rtreeportal.org/
13. Turgay, E.O., Pedersen, T.B., Saygın, Y., Savaş, E., Levi, A.: Disclosure risks of distance preserving data transformations. In: Ludäscher, B., Mamoulis, N. (eds.) SSDBM 2008. LNCS, vol. 5069. Springer, Heidelberg (2008)

# A Novel Approach for Practical Semantic Web Data Management

Giorgio Gianforme[1,*], Roberto De Virgilio[1], Stefano Paolozzi[2],
Pierluigi Del Nostro[1], and Danilo Avola[2]

[1] Università Roma Tre, Rome, Italy
[2] National Research Council, Rome, Italy
giorgio.gianforme@gmail.com, dvr@dia.uniroma3.it,
stefano.paolozzi@irpps.cnr.it, pdn@dia.uniroma3.it,
danilo.avola@irpps.cnr.it

**Abstract.** The growing importance of RDF as a tool for describing information on the Web has risen a number of interesting works regarding the management of RDF documents. The effective management of RDF data has been an increasingly pressing active area of the current research. However current data management solutions for RDF data present the following shortcomings: (i) they often define new query languages difficult to integrate in database applications, with a consequent lack of uniformity among the different approaches, (ii) they present extendibility limitations, and (iii) most of them are considered complicated by the final user. In this paper we propose a feasible and intuitive approach for practical RDF management, providing a logical organization technique for RDF data, making use of a formalism to represent concepts and properties of RDF and RDF(S), and an extension of a DataLog rule system for querying RDF documents.

**Keywords:** Knowledge Management, DataLog, Semantic Web, RDF.

## 1 Introduction

Semantic Web, as firstly proposed by Tim Berners-Lee [2], represents the new generation of Web. The reference model for Semantic Web is the Resource Description Framework or RDF (http://www.w3.org/RDF/). It is a simple logical language which allows the specification of binary properties on Web resources. The simplest possible structure for representing information was chosen in RDF, that is a labeled graph. The assertions, called *triples*, are statements expressing that a *resource*, identified by an URI, is related to another resource or to a value (datatype literal, or XML literal) through a *property* (or *predicate*). In this paper we refer to the term *predicate* as a binary relation between two resources and to the term *property* as a binary relation between a resource and a datatype

[*] Supported in part by Microsoft Research through the European PhD Scholarship Programme.

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 650–655, 2008.

literal. An RDF graph can be viewed as a set of directed edges, commonly represented by triples of form $\langle Subject\ Predicate\ Object \rangle$. Syntactically, this graph can be represented using XML syntax (RDF/XML). The relevant advantage of this approach is that it is very general. Any type of data can be expressed in this format, and it is easy to build tools that manipulate RDF. Furthermore, an increasing amount of data is available on the Web in RDF format.

This data representation, though flexible, faces to serious problems such as data management and query execution. Our purpose is not to extend the RDF data model expressiveness, rather we want to explore ways to improve the organization and query effectiveness on RDF data. Several solutions has been proposed for RDF data management. Some approaches try to extend the expressiveness of RDF and provide ad-hoc query languages (mainly SQL-like) on these extensions [5,6]. However these languages are difficult to integrate in database applications. Other proposals [3,4,9] focus on an effective organization of data in a RDF document, but they face the problem at a physical level for scalability issue, presenting various limitations at logical level (e.g. NULL values, union and joins management, etc.). Moreover most of these solutions present extendibility limitations and provide tools often complex to use.

In this paper we propose a novel approach for a flexible management of RDF documents, based on a logical organization of data. More precisely, the most relevant contributions of our work are:

(i) The creation of a high level description model (called metamodel) to represent the information stored in RDF document, pointing out the implicit semantics of elements through constructs.
(ii) The definition of an intuitive rules-based system, based on an extension of DataLog, to query and manage RDF documents.

The paper is structured as follows. In Section 2 we introduce our metamodel and the rule-based formalism for the management of RDF documents. Finally in Section 3 we sketch concluding remarks and future works.

## 2 Management of RDF Documents

In this section we illustrate our approach to represent RDF and RDF(S) in a compact but expressive form and show how such methodology allows to query and manage RDF documents easily and effectively.

### 2.1 RDF Modeling

Our approach is inspired by works of Atzeni et al. [1,7] that propose a framework for the management of translations between heterogeneous data models in an uniform way. They leverage on the concept of supermodel that allows a high level description of models by means of a generic set of constructs, thus every model can be seen as an instance of the supermodel. The meta description of a model is called metamodel. Following this idea, we propose a simple metamodel where a set of constructs properly represent concepts expressible with RDF and RDF(S): (i) *Class*, to represent an RDF class, (ii) *Property*, to represent an RDF statement that has a primitive type like object, and (iii) *Predicate*, to represent an RDF statement that involves two classes. A difference with respect to the approach of Atzeni et al. is that they consider a well marked distinction between the schema level and the instances while, for our purposes, we need to manage at the same time RDF schemes and instances; for this reason three more constructs are introduced to represent instances, namely: i-Class (resources, with URI), i-Property (with value) and i-Predicate.

More in details, the various constructs have an identifier and a name, are related each other by means of mandatory references and may have properties that specify details of interest. In our case, the *Property* construct has a reference to the *Class* it belongs to and has a *type* property to store the primitive type of the value of its instances; the *Predicate* construct has two *Class* references to represent the subject and the object of a statement and has three boolean properties to represent transitivity, symmetricity and functionality. Constructs of the instance level have a reference to the constructs of schema level they correspond to and inherit from them the same references; moreover, i-property has a *value* property to store the actual value of the property and i-class has a *Name* property to store the URI of a resource. In Figure 1 an UML diagram of our metamodel is represented, where the dashed line divides the RDF(S) constructs (at schema level) from the RDF constructs (at instance level). Enclosed in the dashed box, there is the metamodel core.

As we detail in the following, the approach based on the meta descriptions can be easily extended. When the metamodel is not detailed enough (i.e. expressive enough) new constructs can be added and/or new properties to existing

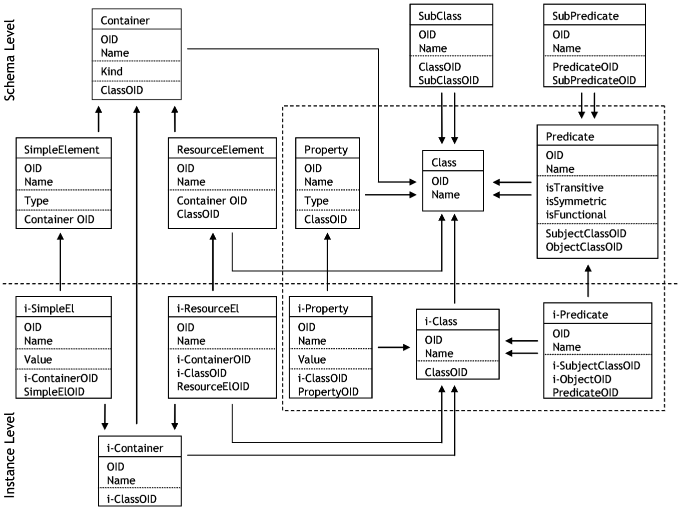


**Fig. 1.** Our metamodel for RDF and RDF(S)

ones. Let us illustrate this point with some examples. Starting from the above mentioned metamodel, if it is also necessary to manage subclasses and subpredicates, it suffices to add two new constructs, namely *SubClass* and *SubPredicate*, each one with two references toward *Class* and *Predicate*. Moreover, three extra constructs are enough for managing RDF collections: *Container*, *SimpleElement* and *ResourceElement*. *Container* represents the entire collection, has a reference to the correspondent class and a property *kind* to denote the type of collection. *SimpleElement* and *ResourceElement* represent elements of a collection that can be, respectively, literals and resources; literal elements have a *type* property while resource elements have a reference to the class and both have a reference to the container to which they belong to. A final remark on our metamodel. Since constructs *SubClass* and *SubPredicate* store meta-information on constructs *Class* and *Predicate*, respectively, they don't have a corresponding construct at the Instance level.

## 2.2 Extending DataLog for RDF Management

The adopted metamodel allows us to exploit rules for querying and maintaining RDF documents that, following [1,7], are expressed referring to metamodel constructs. Such rules are specified in a DataLog extension with OID-invention, by means of Skolem functors. A query or maintenance process is therefore composed by a set of DataLog rules that produce objects (i.e. instances of constructs) belonging to the output, in the first case, to the input, in the latter. The choice of Datalog is essentially due to the high flexibility and to the capacity of expressing recursive queries whereas the lack of recursive primitives in traditional relational languages has represented a major limitation. Then the presence of OID identity, and thus the need for the creation of new objects, each one required to have a new unique object identifier (OID) has brought to the introduction of the notion of OID-invention (e.g. the maintenance of blank nodes).

For instance, let us explain the query process using an example. We consider an RDF document describing family relationships between persons. Each class *Person* has a property, *Name*, and two predicates, *Child* and *Brother*, representing family relationships between persons. There are four instances of the class *Person* (with *URI1*, *URI2*, *URI3* and *URI4* as URI, respectively), each one with a corresponding instance of the property *Name* (with values *Priam*, *Hector*, *Astyanax* and *Paris*, respectively) and linked by three instances of predicate *Child* and one instance of predicate *Brother* representing that *Hector* and *Paris* are sons of *Priam* and brothers and *Astyanax* is son of *Hector*. Through our approach, we can represent this document as depicted in Figure 2, where we omit properties for the sake of simplicity (we don't use them here) and references, that are represented only by arrows. In the figure *URI1*, *URI2*, *URI3* and *URI4* represent the name (as construct property) of the instances of the class *Person*. If a user wants to know just the name of persons that have a child, he needs to specify just two simple rules: first one selects persons that have a child, second one stores the name of such persons; this is done specifying the involved constructs in the body of the rule, relating each other by means of repeated

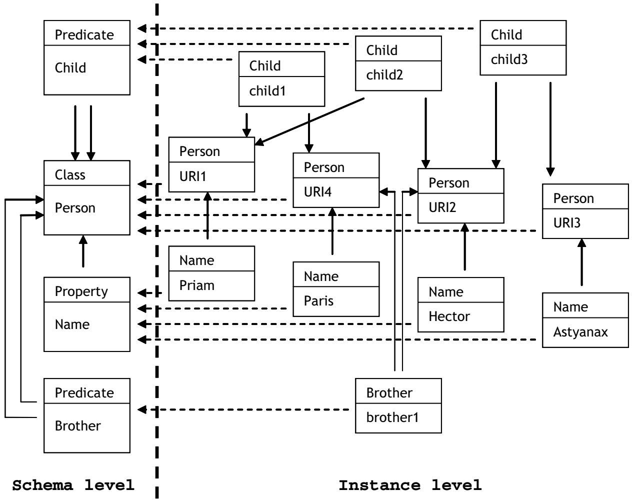


**Fig. 2.** A simple RDF document and its RDF schema

variables for OID's and references, and defining the output of the rule in the head, picking values from the body or creating new ones. The two rules are shown in Figure 3.

We use a non-positional notation for rules, so we indicate just the names of the fields useful for the rule; for example, in the first rule, we use only the field *subjectClassOID* of construct *Child*. Each predicate has an *OID* argument, used for references. Each construct produced by a rule must have a new *OID*, which is generated by means of a Skolem functor, denoted by the # sign in the rules.

Without loss of generality, we assume that our rules satisfy the standard safety requirements [8]: all construct fields in the head have to be defined as a constant, a variable that cannot be undefined (i.e. has to appear somewhere in the body of the rule), or a Skolem term. The same state is for arguments of Skolem terms. Moreover, our DataLog programs are assumed to be coherent with respect to referential constraints. More precisely, if there is a rule that produces a construct $N$ that refers to a construct $N'$, then there is another rule that generates a suitable $N'$ that guarantees the satisfaction of the constraint.

```
Person (OID: #person_0(personOID),
        Name: uri)
←
Person (OID: personOID,
        Name: uri),
Child  (subjectClassOID: personOID);
```

```
Name   (OID: #name_0(nameOID),
        value: value,
        personOID: #person_0(personOID))
←
Name   (OID: nameOID,
        value: value,
        personOID: personOID),
Person (OID: personOID),
Child  (subjectClassOID: personOID);
```

**Fig. 3.** Name of persons that have a child

In the example, the second rule is acceptable because produces constructs *Child* that reference constructs *Person* produced by the first one.

## 3 Conclusions and Future Works

RDF is one of the most representative elements in the Semantic Web. The large amount of information stored as RDF documents in modern WIS has highlighted the importance of methodologies and techniques for managing these documents. In this article we have provided (i) a metamodel approach and (ii) a rule-based system by means of DataLog rules with Skolem functors for the maintenance and querying of RDF documents. With our system it is possible to properly organize RDF documents at a logical level through our metamodel and to easily query them using DataLog rules. Future works will regard the application of the proposed methodology to other fields of interest such as RDF information retrieval and Web services management, scalability and performance issue, and expressiveness comparison of our Datalog system respect to other formalisms.

## References

1. Atzeni, P., Cappellari, P., Bernstein, P.A.: Model-independent schema and data translation. In: Ioannidis, Y., Scholl, M.H., Schmidt, J.W., Matthes, F., Hatzopoulos, M., Böhm, K., Kemper, A., Grust, T., Böhm, C. (eds.) EDBT 2006. LNCS, vol. 3896, pp. 368–385. Springer, Heidelberg (2006)
2. Berners-Lee, T.: Weaving the Web. Orion Business Books (1999)
3. Broekstra, J., Kampman, A., van Harmelen, F.: Sesame: A Generic Architecture for Storing and Querying RDF and RDF Schema. In: Horrocks, I., Hendler, J. (eds.) ISWC 2002. LNCS, vol. 2342. Springer, Heidelberg (2002)
4. Chong, E.I., Das, S., Eadon, G., Srinivasan, J.: An Efficient SQL-based RDF Querying Scheme. In: Proc. of the 31st International Conference on Very Large Data Bases (VLDB 2005), Trondheim, Norway (2005)
5. Furche, T., Linse, B., Bry, F., Plexousakis, D., Gottlob, G.: RDF Querying: Language Constructs and Evaluation Methods Compared. In: Proc. of Int. Summer School on Reasoning Web, Lisbon, Portugal (2006)
6. Polleres, A.: From sparql to rules (and back). In: Proc. of the Int. Conference of World Wide Web (WWW 2007), Banff, Canada (2007)
7. Torlone, R., Atzeni, P.: A unified framework for data translation over the Web. In: Proc. of the 2nd Int. Conf. of Web Information System (WISE 2001), Japan (2001)
8. Ullman, J.D., Widom, J.: A First Course in Database Systems. Prentice-Hall, Englewood Cliffs (1997)
9. Wilkinson, K., Sayers, C., Kuno, H., Reynolds, D.: Efficient RDF Storage and Retrieval in Jena2. In: Proc. of the first International Workshop on Semantic Web and Databases (SWDB 2003), Berlin, Germany (2003)

# Distance-Based Classification in OWL Ontologies

Claudia d'Amato, Nicola Fanizzi, and Floriana Esposito

Department of Computer Science, University of Bari
{claudia.damato,fanizzi,esposito}@di.uniba.it

**Abstract.** We propose inductive distance-based methods for instance classification and retrieval in ontologies. Casting retrieval as a classification problem with the goal of assessing the individual class-memberships w.r.t. the query concepts, we propose an extension of the *k-Nearest Neighbor* algorithm for OWL ontologies based on an *epistemic* distance measure. The procedure can classify the individuals w.r.t. the known concepts but it can also be used to retrieve individuals belonging to query concepts. Experimentally we show that the behavior of the classifier is comparable with the one of a standard reasoner. Moreover we show that new knowledge (not logically derivable) is induced. It can be suggested to the knowledge engineer for validation, during the ontology population task.

## 1 Introduction

Classification for retrieving resources in a knowledge base (KB) is generally performed through logical approaches that may fail in distributed settings, such as the Semantic Web (SW) context, since they are exposed to inconsistency. Another problem is related to the inherent incompleteness of the KBs in the SW applications, where new resources (web docs or services) are likely to be made available along the time. Statistical methods may be suitable for distributed KBs since they can be often efficient and noise-tolerant. An inductive distance-based method for *concept retrieval* may also suggest new assertions which could not be logically derived, providing also a measure of their likelihood which may help dealing with the uncertainty caused by the incompleteness of the KBs. The time-consuming ontology population task can be facilitated since the knowledge engineer would only have to validate the suggested assertions [1].

Retrieval can be cast as a classification problem, i.e. assessing the class-membership of the individuals in the KB w.r.t. query concepts. Similar individuals should likely belong to similar concepts. Moving from such an idea, an instance-based framework for retrieving resources contained in OWL KBs has been devised. Differently from logic-based approaches to (approximate) instance retrieval [4], we propose an extension of the *Nearest Neighbor* (NN) search to the standard representations for ontologies. Our procedure retrieves individuals belonging to query concepts, by analogy with other training instances, based on the classification of the nearest ones in terms of a dissimilarity measure. Extending the NN search to expressive representations founded in Description Logics (DLs) requires suitable metrics. NN search is generally devised for settings where classes are assumed to be disjoint, which is unlikely in the SW context where an individual can be mapped to a hierarchy of concepts. Furthermore, the DL reasoners make the *Open World Assumption* (OWA), differently from the typical (deductive) database engines working with the *Closed World Assumption* (CWA).

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 656–661, 2008.

For our purposes, fully semantic metrics [3] are adopted. These language-independent measures assess the dissimilarity of two individuals by comparing them on the grounds of their behavior w.r.t. a committee of features (concepts) that are defined in the KB or that are be generated to this purpose. All the features have the same importance in determining the dissimilarity. However, it may well be that some features have a larger discriminating power w.r.t. the others. In this case, they should be more relevant in determining the dissimilarity value. We propose an extension of former measures, where each feature of the committee is weighted on the grounds of the amount of information that it conveys. This weight is then determined as an *entropic* measure.

The measure has been integrated in the NN procedure [2] and the classification of resources (individuals) w.r.t. a query concept has been performed through a voting procedure weighted by the neighbors' similarity (Sect. 2). The resulting system allowed for an experimentation of the method on performing instance classification with a number ontologies drawn from public repositories (Sect. 4). Its predictions were compared to assertions that were logically derived by a deductive reasoner. The experiments shows that the classification results are comparable (although slightly less complete) and also that the classifier is able to induce new knowledge that is not logically derivable.

## 2 Resource Retrieval as Nearest Neighbor Search

In the following, we assume that concept descriptions are defined in terms of a DL language that can be mapped to OWL-DL. A *knowledge base* $\mathcal{K} = \langle \mathcal{T}, \mathcal{A} \rangle$ contains a *TBox* $\mathcal{T}$ and an *ABox* $\mathcal{A}$. $\mathcal{T}$ is a set of (equivalence or also inclusion) axioms that define concepts. $\mathcal{A}$ contains factual assertions concerning the resources, also known as *individuals*. The set of the individuals occurring in $\mathcal{A}$ will be denoted with $\mathsf{Ind}(\mathcal{A})$. As regards the inference services, our procedure may require performing *instance-checking*, namely determining whether an individual, say $a$, belongs to a concept extension, i.e. whether $C(a)$ holds for a certain concept $C$. Because of the OWA, it can happen that a reasoner may be unable to give a positive or negative answer to a class-membership query.

Query answering boils down to determining whether a resource belongs to a concept. Here, an alternative inductive method is proposed. It consists in casting the query answering problem as determining the correct classification for a query individual. The method is grounded on the NN search. It could be able to provide an answer when it may not be inferred by deduction. Moreover, it may also provide a measure of the likelihood of its answer. The basic idea of the NN search is to find the most similar individuals w.r.t. the one that has to be classified. Let $x_q$ be the query instance whose class-membership has to be determined. Using a dissimilarity measure, the set of the $k$ nearest (pre-classified) training instances w.r.t. $x_q$ is selected: $NN(x_q) = \{x_i \mid i = 1, \ldots, k\}$. The objective is to induce an approximation for a discrete-valued target hypothesis function $h : \mathit{IS} \mapsto V$ from a space of instances *IS* to a set of values $V = \{v_1, \ldots, v_s\}$ standing for the classes (concepts) that have to be predicted. In its simplest setting, the $k$-NN algorithm approximates $h$ for classifying $x_q$ on the grounds of the (weighted) value that $h$ is known to assume for the training instances in $NN(x_q)$ as follows:

$$\hat{h}(x_q) := \operatorname*{argmax}_{v \in V} \sum_{i=1}^{k} w_i \delta(v, h(x_i)) \tag{1}$$

where $\delta$ returns 1 in case of matching arguments and 0 otherwise, and, given a dissimilarity measure $d$, the weights are determined by $w_i = 1/d(x_i, x_q)$.

This setting assigns a value to the query instance which stands for one in a set of pairwise disjoint concepts (corresponding to the value set $V$). In a multi-relational setting this assumption cannot be made in general. An individual may be an instance of more than one concept. Moreover, to deal with the OWA, the absence of information on whether a training instance $x$ belongs to the extension of the query concept $Q$ should not be interpreted negatively, as in the standard settings which adopt the CWA. Rather, it should count as neutral (uncertain) information. In order to solve this problems, the multi-class problem is transformed into a ternary one and another value set $V = \{+1, -1, 0\}$ is adopted, where the three values denote, respectively, membership, non-membership, and uncertainty. Specifically, the task can be cast as follows: given a query concept $Q$, determine the membership of an instance $x_q$ through the NN procedure (see Eq. 1) where $V = \{-1, 0, +1\}$ and the hypothesis function values for the training instances are determined by the entailment of the corresponding assertions from the knowledge base, as follows:

$$h_Q(x) = \begin{cases} +1 & \mathcal{K} \models Q(x) \\ -1 & \mathcal{K} \models \neg Q(x) \\ \;\;0 & \textit{otherwise} \end{cases}$$

Note that, being the procedure based on a majority vote of the individuals in the neighborhood, it is less error-prone in case of noise in the data (e.g. incorrect assertions) w.r.t. a purely logic deductive procedure. Therefore, it may be able to give a classification even in case of inconsistent knowledge bases. However, the classification result is not guaranteed to be deductively valid. Indeed, inductive inference naturally yields a certain degree of uncertainty. In order to measure the likelihood of the decision made by the procedure, the quantity that determined the decision should be normalized by dividing it by the sum of such arguments over the (three) possible values:

$$l(\mathit{class}(x_q) = v | NN(x_q, k)) = \frac{\sum_{i=1}^{k} w_i \cdot \delta(v, h_Q(x_i))}{\sum_{v' \in V} \sum_{i=1}^{k} w_i \cdot \delta(v', h_Q(x_i))} \tag{2}$$

Hence the likelihood of the assertion $Q(x_q)$ corresponds to the case when $v = +1$.

## 3 A Family of Epistemic Metrics for Individuals

For the NN procedure, we intend to exploit a family of new measures for DL representations that totally depend on semantic aspects of the individuals in the KB. They are based on the idea that, on a semantic level, similar individuals should behave similarly w.r.t. the same concepts. The rationale is to compare individuals on the grounds of their semantics w.r.t. a collection of concept descriptions, say $\mathsf{F} = \{F_1, F_2, \ldots, F_m\}$, which stands as a group of discriminating *features* expressed in the OWL-DL sub-language taken into account. They are formally defined as follows [3]:

**Definition 3.1 (family of measures).** *Let $\mathcal{K} = \langle \mathcal{T}, \mathcal{A} \rangle$ be a knowledge base. Given the set of concept descriptions $\mathsf{F} = \{F_1, F_2, \ldots, F_m\}$, a family of functions $\{d_p^{\mathsf{F}}\}_{p \in \mathbb{N}}$ with $d_p^{\mathsf{F}} : \mathsf{Ind}(\mathcal{A}) \times \mathsf{Ind}(\mathcal{A}) \mapsto [0, 1]$ is defined as follows:*

$$\forall a, b \in \mathsf{Ind}(\mathcal{A}) \qquad d_p^{\mathsf{F}}(a,b) := \frac{L_p(\pi(a),\pi(b))}{m} = \frac{1}{m}\left(\sum_{i=1}^{m} \mid \pi_i(a) - \pi_i(b) \mid^p\right)^{\frac{1}{p}}$$

*where* $p > 0$*, the weights* $w_i \in [0,1]$*,* $1 \leq i \leq m$*, and the* projection function $\pi_i$ *is:*

$$\forall a \in \mathsf{Ind}(\mathcal{A}) \quad \pi_i(a) = \begin{cases} 1 & \mathcal{K} \models F_i(a) \\ 0 & \mathcal{K} \models \neg F_i(a) \\ 1/2 & otherwise \end{cases}$$

Note that in the measure definition, the features have all the same weights. However, each feature could have a different discriminating power. In order to take into account such an aspect, a weight for each feature is introduced. It is determined by exploiting the quantity of information conveyed by the feature, namely by measuring the feature entropy as: $H(F) = -\left(P_F \log(P_F) + P_{\neg F}\log(P_{\neg F}) + P_U \log(P_U)\right)$ where $P_F$ represents the probability of the feature $F$ and is computed as: $P_F = |\mathsf{retrieval}(F)|/|\mathsf{Ind}(\mathcal{A})|$.

## 4 Experiments

The NN procedure integrated with the new distance has been tested for solving a number of retrieval problems. To this purpose, we selected several OWL ontologies, summarized in Tab. 1, available on the web. For each ontology, 20 queries were randomly generated by composition (conjunction and/or disjunction) of (2 through 8) concepts or restrictions of object and data-properties. The performance of the inductive method was evaluated by comparing its responses to those returned by PELLET reasoner. We selected limited training sets ($\mathsf{TrSet}$) that amount to only 4% of the individuals occurring in each ontology. The parameter $k$ was set to $\sqrt{|\mathsf{TrSet}|}$. The simpler distances ($d_1^{\mathsf{F}}$) were employed from the family (weighted on the feature entropy), using all the concepts in the KB for determining the set $\mathsf{F}$. We performed two experiments: one aiming at a three-way classification and the other forcing the response to attribute each test instance to either the target class or to its negation. Initially the standard measures precision, recall, F$_1$-measure were employed to evaluate the system performance. However, due to the OWA, several cases were observed when, it could not be (deductively) ascertained whether a resource was relevant or not for a given query. Hence, we have introduced the following further evaluation indices: 1) *match rate*: number of individuals that got exactly the same classification ($v \in V$) by both the inductive and the deductive classifier w.r.t. the overall number of individuals ($v$ vs. $v$); 2) *omission error rate*: amount

**Table 1.** Facts concerning the ontologies employed in the experiments

| Ontology | DL language | #concepts | #object prop. | #data prop. | #individuals |
|---|---|---|---|---|---|
| SWM | $\mathcal{ALCOF}(D)$ | 19 | 9 | 1 | 115 |
| BIOPAX | $\mathcal{ALCHF}(D)$ | 28 | 19 | 30 | 323 |
| LUBM | $\mathcal{ALR}^+\mathcal{HI}(D)$ | 43 | 7 | 25 | 555 |
| NTN | $\mathcal{SHIF}(D)$ | 47 | 27 | 8 | 676 |
| SWSD | $\mathcal{ALCH}$ | 258 | 25 | 0 | 732 |
| FINANCIAL | $\mathcal{ALCIF}$ | 60 | 17 | 0 | 1000 |

**Table 2.** Average results (percentages) of the 3-way classification experiment

| | precision | recall | F-measure | match | commission | omission | induction |
|---|---|---|---|---|---|---|---|
| SWM | 99.0 | 75.8 | 79.5 | 97.5 | 0.0 | 2.2 | 0.3 |
| BIOPAX | 99.9 | 97.3 | 98,2 | 99.9 | 0.1 | 0.0 | 0.0 |
| LUBM | 100.0 | 81.6 | 85.0 | 99.5 | 0.0 | 0.5 | 0.0 |
| NTN | 97.0 | 40.1 | 45.1 | 97.5 | 0.6 | 1.3 | 0.6 |
| SWSD | 94.1 | 38.4 | 46.5 | 98.0 | 0.0 | 1.9 | 0.1 |
| FINANCIAL | 99.8 | 95.0 | 96.6 | 99.7 | 0.0 | 0.0 | 0.2 |

**Table 3.** Average results (percentages) of the 2-way classification experiment

| | precision | recall | F-measure | match | commission | omission | induction |
|---|---|---|---|---|---|---|---|
| SWM | 60.3 | 100.0 | 75.2 | 59.9 | 0.0 | 0.0 | 40.1 |
| BIOPAX | 92.2 | 58.4 | 71.5 | 87.1 | 12.9 | 0.0 | 0.0 |
| LUBM | 63.1 | 100.0 | 65.3 | 63.0 | 0.0 | 0.0 | 37.0 |
| NTN | 44.9 | 100.0 | 48.6 | 40.2 | 0.3 | 0.0 | 59.5 |
| SWSD | 6.3 | 100.0 | 7.5 | 6.2 | 0.0 | 0.0 | 93.8 |
| FINANCIAL | 86.8 | 72.6 | 71.8 | 91.6 | 6.2 | 0.0 | 2.2 |

of individuals for which inductive method returns 0 while they were actually relevant according to the reasoner (0 vs. $\pm 1$); 3) *commission error rate*: number of individuals found to be relevant to the query concept, while they actually belong to its negation or vice-versa ($+1$ vs. $-1$ or $-1$ vs. $+1$); 4) *induction rate*: amount of individuals found to be relevant to the query concept or to its negation, while either case is not logically derivable from the KB ($\pm 1$ vs. 0). For each KB, we report the average values (percentages) obtained over the 20 query concepts randomly generated. The outcomes of the three-way classification experiments are reported in Tab. 2. Note that precision and recall are generally quite good for all ontologies but SWSD, where especially recall is significantly lower. SWSD turned out to be more difficult (also in terms of precision) for two reasons: a very limited number of individuals per concept was available and the number of different concepts is larger w.r.t. the other KBs. For the other ontologies values are much higher, as testified also by the F-measure values. Moreover, the results in terms of precision are more stable than those for recall as proved by the limited variance observed, whereas single queries happened to turn out quite difficult as regards the correctness of the answer. The reason for precision being generally higher is probably due to the OWA. Indeed, in many cases it was observed that the NN procedure deemed some individuals as relevant for the query issued while the reasoner was not able to assess this relevance and this was computed as a mistake while it may likely turn out to be a correct inference when judged by a human agent. It is also important to note that, in each experiment, the commission error was quite low or absent. This means that the inductive search procedure did not make critical mistakes. Also the omission error rate was generally quite low, yet more frequent than the previous type of error.

Tab. 3 reports the outcomes of the 2-way experiment where omission errors were ruled out. As expected a dramatic increase of inductive assertions was observed since the system is not trying to compare its inferences to deductive ones but rather it is trying to suggest potential class-memberships. However the observed commission error rates were still quite low which shows that the system was really forcing an answer in the unknown cases and this answer is likely to be correct, at least for those individuals for which the classification can be logically derived from the knowledge base. Conversely, the reported recall rates are higher than with the 3-way setting. The noteworthy low precision and match rate for SWSD (and, partially, for NTN), can be explained by the fact that this ontology has been artificially populated with very few (often 2) individuals per concept, which is harmful for instance-based methods like nearest-neighbor search, especially in a two-classification setting. However less precision is often compensated by high induction rates. The usage of all concepts for the set $\mathsf{F}$ of $d_1^{\mathsf{F}}$ made the measure quite accurate, which is the reason why the procedure resulted quite conservative as regards inducing new assertions. In many cases, it matched rather faithfully the reasoner decisions. From the retrieval point of view, the cases of induction are interesting because they suggest new assertions which cannot be logically derived by using a deductive reasoner yet they might be used to complete a knowledge base [1], e.g. after being validated by an ontology engineer. For each candidate new assertion, Eq. 2 may be employed to assess the likelihood and hence decide on its inclusion. If we compare these outcomes with those reported in other works on instance retrieval and inductive classification [2], where the highest average match rate observed was around 80%, we find a significant increase of the performance due to the accuracy of the new measure.

## 5 Conclusions

This paper investigated the application of a distance-based classification method for KBs represented in OWL. We employed an extended family of dissimilarity measures based on feature committees [3] taking into account the amount of information conveyed by each feature based on an estimate of its entropy. The measures were integrated in a distance-based search procedure that have been exploited for the task of approximate instance retrieval. The experiments made showed that the method is quite effective and can be applied to any domain.

## References

[1] Baader, F., Ganter, B., Sertkaya, B., Sattler, U.: Completing description logic knowledge bases using formal concept analysis. In: Veloso, M. (ed.) Proc. of IJCAI (2007)

[2] d'Amato, C., Fanizzi, N., Esposito, F.: Reasoning by analogy in description logics through instance-based learning. In: Proc. of SWAP 2006 Workshop, vol. 201. CEUR (2006)

[3] Fanizzi, N., d'Amato, C., Esposito, F.: Induction of optimal semi-distances for individuals based on feature sets. In: Working Notes of DL 2007 Workshop, vol. 250. CEUR (2007)

[4] Möller, R., Haarslev, V., Wessel, M.: On the scalability of description logic instance retrieval. vol. 189. CEUR (2006)

# Kernel Methods for Graphs: A Comprehensive Approach

Francesco Camastra and Alfredo Petrosino

Department of Applied Science, University of Naples Parthenope
Centro Direzionale Isola C4, 80143 Naples, Italy
{francesco.camastra, alfredo.petrosino}@uniparthenope.it

**Abstract.** The development of learning algorithms for structured data, i.e. data that cannot be represented by numerical vectors, is a relevant challenge in machine learning. Kernel Methods, which is a leading machine learning technology for vectorial data, recently tackled the structured data. In this paper we focus our attention on Kernel Methods that face up to data that can be represented by means of graphs, by providing an in-depth review through a comprehensive approach to the research hints and the main open problems in this area of research.

## 1 Introduction

Machine learning methods generally work on fixed-length vectorial data. However, in real world applications we can deal with data that cannot be represented by vectorial data but by structures. These types of data are called *structured data*. Examples of structured data are molecular formulae and images when represented by hierarchical structures. For instance, an image can be segmented into regions than may be recursively segmented into several sub-regions resulting in a hierarchical tree structure. Several supervised and unsupervised neural network models for structured data, such as trees and graphs, have been proposed [1,2]. Recently Kernel Methods [3] for structured data [4] have been developed. In this manuscript we will focus our attention on Kernel Methods that tackle structured data, i.e. data that can be represented by means of graphs. We shall provide a comprehensive framework that allows to apply the main Kernel Methods for graphs to various applications, like bioinformatics (e.g. molecular data mining, chemogenomics, etc.) and computer vision (e.g. 3D object structural recognition, spatio-temporal data mining, etc.). The paper is organized as follows: In Section 2 main facts on labelled graphs are recalled; the general issues of Kernel Methods are reviewed in Section 3; Product Graph Kernel is presented in Section 4; Marginalized Kernels and its extensions are described in Section 5; Cyclic Pattern Kernels for Graphs are discussed in Section 6, respectively; finally some research hints and the main open problems are discussed in Sections 7 and 8.

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 662–669, 2008.

## 2 Labelled Graphs

A *graph* $G$ is described by a finite set of vertices $V = \{v_1, \ldots, v_n\}$, a finite set of edges $E = \{e_1, \ldots, e_p\} \subseteq V \times V$ and a function $\Psi : E \rightarrow V \times V$. A *labelled graph* is a tuple $(V, E, \Lambda, \lambda)$ where $\Lambda = (l_1, \ldots, l_r)$ is a set of labels and $\lambda : \Lambda \rightarrow V \cup E$ is a function which assigns a label to each edge and vertex. In a *directed graph* the function $\Psi$ maps each edge to the tuple composed of its initial and terminal node. A *loop* is an edge $e$ in a directed graph such that $\psi(e) = (v, v)$. Two edges $e_1$ and $e_2$ are *parallel* iff $\Psi(e_1) = \Psi(e_2)$. A *walk* is a sequence of vertices $v_i \in V$ and $e_i \in E$ with $w = v_1, e_1, v_2, e_2, \ldots, v_n, e_n, v_{n+1}$ and $\Psi(e_i) = (v_i, v_{i+1})$. The *length* of the walk is given by the number $n$ of edges in the sequence. A *path* is a walk such that $v_i \neq v_j \iff i \neq j$ and $e_i \neq e_j \iff i \neq j$. A *cycle* is a path with an edge $e_{n+1}$ such that $\Psi(e_{n+1}) = (v_{n+1}, v_n)$. A graph is *connected* if a path exists between any couple of vertices. To describe the neighborhood of a vertex $v$ in a graph $G$ we need some further definitions. The set $\delta^+(v) = \{e \in E | \Psi(e) = (v, u)\}$ denotes the set whose elements come out from the edge $v$. Its cardinality $|\delta^+(v)|$ is called the *outdegree* of the vertex $v$. In similar way, the set $\delta^-(v) = \{e \in E | \Psi(e) = (u, v)\}$ defines the set whose elements come into the vertex $v$. The cardinality $|\delta^-(v)|$ is called the *indegree* of the vertex $v$. The maximal and the minimal indegree are $\Delta^-(G) = \max\{|\delta^-(v)|, v \in V\}$ and $\Delta^-(G) = \max\{|\delta^+(v)|, v \in V\}$, respectively. To represent a graph we use the *adjacency matrix* $E$, where $E_{ij}$ of the matrix provides the number of edges between vertex $v_i$ and $v_j$. The element $E^n_{ij}$ of the $n^{th}$ power of the adjacency matrix $E^n$ gives the number of walks of length $n$ from the vertex $v_i$ and $v_j$. Now we introduce the concept of *labelled subgraph.* Let $G = (V, E, \Lambda, \lambda)$ and $G' = (V', E', \Lambda, \lambda')$ be two labelled graphs. $G'$ is a labelled subgraph of $G$ if $V' \subseteq V$, $E' \subseteq E$ and $\lambda'(x) = \lambda(x)$ for all $x \in V' \cup E'$. A maximal connected subgraph of the graph $G$ is called a *connected component* of $G$. A vertex $v$ of a graph $G$ is called a *cut vertex* if the subgraph $G'$ obtained from $G$ removing $v$ and all edges that comes into or comes out the vertex $v$ has more connected components than $G$. A graph is *biconnected* if has no cut vertex. A biconnected maximal subgraph of a graph $G$ is called a *biconnected component* of a graph $G$. The biconnected components of a graph are pairwise edge disjoint inducing a partition on the set $E$ of the graph $G$. The partition implies that two edges are equivalent iff they belong to a common cycle. Therefore an edge of a graph belongs to a cycle iff the biconnected components of the graph have more than one edge. Edges that do not belong to cycles are called *bridges.* The set of the bridges of a graph $G$ forms a subgraph of $G$ denoted by $B(G)$.

## 3 Kernel Methods

Firstly we recall the definition of *Mercer kernel* (or *positive definite kernel*) [5].

**Definition 1.** *Let $X$ be a nonempty set. A function $K : X \times X \rightarrow \mathbb{R}$ is called a* Mercer kernel *if it is* symmetric *and $\sum_{j=1}^n \sum_{k=1}^n c_j c_k K(x_j, x_k) \geq 0$ for all $n \geq 2$, $\{x_1, \ldots, x_n\} \subseteq X$ and $\{c_1, \ldots, c_n\} \subseteq \mathbb{R}$. Each Mercer kernel $K$ can* *be represented as:* $K(x,y) = \langle \Phi(x), \Phi(y) \rangle$ *where* $\langle \cdot, \cdot \rangle$ *is the inner product and* $\Phi : X \to \mathcal{F}$. $\mathcal{F}$ *is called the* Feature Space.

The simplest example of the Mercer kernel is the *inner product* $K(\boldsymbol{x}, \boldsymbol{y}) = \langle \boldsymbol{x}, \boldsymbol{y} \rangle$. A popular Mercer kernel is the *Gaussian* $K(\boldsymbol{x}, \boldsymbol{y}) = e^{-\frac{\|\boldsymbol{x}-\boldsymbol{y}\|^2}{\sigma^2}}$, where $\sigma \in \mathbb{R}$. An important Mercer kernel for the rest of the paper is the *Dirac kernel* $\delta(x,y)$, that is 1 if $x = y$ and 0 otherwise. Mercer kernels can also be defined on structured data. An example of Mercer kernel on structured data is the *intersection kernel* between sets [3].

**Definition 2.** *Let* $I_A$ *be the indicator function of a measurable set* $A$ *(i.e* $I_A(x) = 1$ *if* $x \in A$, $I_A(x) = 0$ *otherwise). The indicator function defines the measure* $\mu$ *of the set* $A$ *such as* $\mu(A) = \mu(I_A)$. *Given two sets* $A_1$ *and* $A_2$ *the* intersection kernel $K_\cap$ *is defined by:* $K_\cap(A_1, A_2) = \mu(A_1 \cap A_2)$.

Kernel Methods [3] are algorithms that implicitly perform, by replacing the inner product by an appropriate Mercer Kernel, a nonlinear mapping of the input data to a high dimensional Feature Space. The most popular Kernel Method is the Support Vector Machine (SVM) [3]. In Kernel Methods the data only occur under the form of an appropriate Mercer kernel input. Therefore they can also be used for structured data whenever it is possible to define a Mercer kernel on them. Hence the research on Kernel Methods for structured data results in designing appropriate Mercer kernels on them. The rest of the paper is focused on the description of the main Mercer kernels that it is possible to define on graphs.

## 4 Product Graph Kernel

In this section we review the *Product Graph Kernel* [6] that is based on the idea of counting the number of walks in product graphs. Product graphs can be defined in several ways. In this context we will only consider the direct product of two labelled graphs, that is defined as follows.

**Definition 3.** *The direct product of two labelled graphs* $G_1 = (V_1, E_1, \Lambda, \lambda_1)$ *and* $G_2 = (V_2, E_2, \Lambda, \lambda_2)$ *is denoted by* $G_1 \times G_2$. *The vertex set of the direct product is defined as:* $V(G_1 \times G_2) = \{(v_1, v_2) \in V_1 \times V_2 : \delta(\lambda_1(v_1), \lambda_2(v_2)) = 1\}$. *The edge set is defined as:*

$$\begin{aligned} E(G_1 \times G_2) = \{(e_1, e_2) \in E_1 \times E_2 : \exists (u_1, u_2), (v_1, v_2) \in V(G_1 \times G_2) \\ \wedge \Psi_1(e_1) = (u_1, v_1) \wedge \Psi_2(e_2) = (u_2, v_2) \\ \wedge (\delta(\lambda_1(e_1), \lambda_2(e_2)) = 1)\}, \end{aligned}$$

*where* $\delta(\cdot)$ *is the Dirac Kernel. Given an edge* $(e_1, e_2) \in E(G_1 \times G_2)$ *with* $\Psi_1(e_1) = (u_1, v_1)$ *and* $\Psi_2(e_2) = (u_2, v_2)$ *the value of* $\Psi_{G_1 \times G_2}$ *is:*

$$\Psi_{G_1 \times G_2}((e_1, e_2)) = ((u_1, u_2), (v_1, v_2)).$$

*The label of the vertices and edges in graph* $G_1 \times G_2$ *correspond to the labels in the factors. The graphs* $G_1$, $G_2$ *are called the factors of graph* $G_1 \times G_2$.

Now we can define the *product graph kernel.*

**Definition 4.** *Let $G_1$, $G_2$ be two labelled graphs, let $A_\times = A(G_1 \times G_2)$ be the adjacency matrix of their direct product and let $V_\times = V(G_1 \times G_2)$ be the vertex set of the direct product. With a sequence of weights $\lambda = \lambda_0, \lambda_1, \ldots$ (where $\lambda_i \in \mathbb{R}^+$) the product graph kernel is defined as $K_\times(G_1, G_2) = \sum_{i,j=1}^{|V_\times|} \left[\sum_{n=0}^{\infty} \lambda_n A_\times^n\right]_{ij}$.*

The product graph kernel, that is a Mercer kernel [6], in conjunction with Gaussian Processes [7] has been used to determinate an optimal strategy for the play Tetris [8]. Now we show how the product graph kernel can be computed. Firstly, we discuss the exponential setting [9], i.e. we set $\lambda_i = \frac{\beta_i}{i!}$. The exponential of the square matrix $A$ is defined as: $e^{\beta A} = \lim_{n\to\infty} \sum_{i=0}^{n} \frac{(\beta A)^i}{i!}$ with $\frac{\beta^0}{0!} = 1$ and $A^0 = \mathbb{I}$, where $\mathbb{I}$ denotes the identity matrix and $\beta$ is a constant that must be set up. If we represent the matrix $A$ in terms of a diagonal matrix $D$, i.e. $A = T^{-1}DT$, we obtain: $K_\times(G_1, G_2) = \sum_{i,j=1}^{|V_\times|} \left[T^{-1} e^{\beta D_\times} T\right]_{ij}$. The product direct kernel can also be computed by the geometric series. The geometric series $\sum_{i=0}^{\infty} \gamma^i$ converges iff $|\gamma| < 1$ and its limit is given by: $\sum_{i=1}^{\infty} \gamma^i = \frac{1}{1-\gamma}$. In similar way we can define the geometric series of a matrix $A$ as follows: $\sum_{i=0}^{\infty} \gamma^i A^i$. If the matrix $A$ is symmetric and $\gamma < 1$, the limit of the geometric series is given by: $\sum_{i=0}^{\infty} \gamma^i A^i = (\mathbb{I} - A)^{-1}$. Hence the product graph kernel by geometrical series is: $K_\times(G_1, G_2) = \sum_{i,j=1}^{|V_\times|} [(\mathbb{I} - A_\times)^{-1}]_{ij}$.

## 5 Marginalized Graph Kernel

In this section we will describe the *marginalized graph kernel* [10]. Given two labelled graphs $G_1$ and $G_2$, the idea behind marginalized graph kernel consists in comparing the label sequences generated by the two synchronized random walks of both graphs. Let $G = (V, E, \Lambda, \lambda)$ be a labelled graph, where $V = (v_1, \ldots, v_n)$ and $E = (e_1, \ldots, e_p)$ are the sets of vertices and edges, respectively. A random walk on the graph produces a *path*, that depends on the initial probability $P_s(v)$ distribution on $V$, a transition probability $P(v_i|v_{i-1})$ from the vertex $v_{i-1}$ to the vertex $v_i$, and the probability $P_q(v_\ell)$ of stopping the random walk at vertex $\ell$. If we do not use prior knowledge, $P_s(v)$ is a uniform distribution, $P(v_i|v_{i-1})$ is a uniform distribution over all adjacent vertices of the actual one, and $P_q(v_\ell)$ is a constant equal to $\frac{1}{n}$. From a given path obtained from the random walk, we can define the label sequence as an alternative sequence of vertex and edge labels, i.e. $h = (v_1, e_{1,2}, v_2, \ldots, v_{\ell-1}, e_{\ell-1,\ell}, v_\ell)$. The marginalized graph kernel between two labelled graphs $G_1 = (V_1, E_1, \Lambda, \lambda_1)$ and $G_2 = (V_2, E_2, \Lambda, \lambda_2)$ is defined as the expectation of the kernel between sequences $K(h_1, h_2)$ over all possible paths of all lengths:

$$K(G_1, G_2) = \sum_h \sum_{h'} K_\lambda(l(h_1), l(h_2)) p_1(h_1) p_2(h_2), \tag{1}$$

where $p_1$ and $p_2$ are probability distributions on the set of finite-length sequences of the vertices of the graphs $G_1$ and $G_2$, respectively and $K_\lambda(\cdot)$ is an appropriate kernel between label sequences, i.e. $K_\lambda : \Lambda^\star \times \Lambda^\star \to \mathbb{R}$, where $\Lambda^\star$ denotes the set of finite-length sequence of labels of $\Lambda$. The marginalized graph kernel can be interpreted as a marginalized kernel [11], motivating the name and guaranteeing that is a Mercer kernel. Kashima et al. [10] have investigated a particular case of the marginalized graph kernel when the $K_\lambda$ is the Dirac kernel and the probability $p$ of the equation (1) can be expressed as:

$$p(v_1, \ldots, v_n) = p_s(v_1) \prod_{i=2}^{n} p_t(v_i | v_{i-1}). \tag{2}$$

To guarantee that $p$ is a probability, i.e. $\sum_{v \in V} p(v) = 1$, it is necessary that some $p_s$ and $p_t$ fulfill some constraints. If we choose

- a stopping probability $p_q$ (with $0 < p_q(v) < 1$) for each vertex $v$,
- an initial probability distribution $p_0$ (such that $\sum_{v \in V} p_0(v) = 1$),
- a transition probability $p_a$ on $V \times V$ (with $\sum_{v \in V} p_a(u|v) = 1$) such that if $p_a(v|u) > 0$ then $(u, v) \in E$, i.e. $p_a(v|u)$ is positive only when an edge exists,
- and for any $u, v \in V^2$ $p_s(v) = p_0(v)p_q(v)$ and $p_t(u|v) = \frac{1-p_q(v)}{p_q(v)} p_a(u|v)p_q(u)$;

then $p$, defined in (2), is a probability distribution that corresponds to a random walk on the graph $G$ with an initial, transition and stopping probability defined by $p_0$, $p_a$ and $p_s$, respectively.

Marginalized graph kernel has two main drawbacks. The former consists in its computational burden. Therefore its usage is not advisable when the data set contains several thousands of graphs, as it generally happens in bioinformatics applications The latter is represented by the strategy implemented in the marginalized graph kernel. The kernel is based on the search of common paths between graphs which may be too simple to detect common patterns of interest between graphs. To overcome these limitations, extensions of marginalized graph kernel [12] have been proposed. They increase the degree of the specificity of the label vertices of the graph, taking into account contextual information about the vertices.

## 6 Cyclic Pattern Kernels for Graphs

In this section we describe the *cyclic pattern kernels for graphs* (CPKs) [13], that are based on the intersection kernel between sets. To use the intersection kernel on graphs, each graph $G$ is represented by a set of cycles and tree patterns of $G$. Let $G = (V, E, \Sigma, \lambda)$ be a graph and $C = \{v_0, v_1\}, \{v_1, v_2\}, \ldots, \{v_{k-1}, v_0\}$ be a sequence of $k-1$ edges that forms a cycle in $G$. The canonical representation of $C$ is defined as follows. If we denote by $\rho(s)$ the set of cyclic permutations of a sequence $s$ and its reverse, the *canonical representation* $\pi(C)$ [14] is defined by:

$$\pi(C) = \min(\sigma(w) : w \in \rho(v_0 v_1 \ldots v_{k-1}))$$

where for $w = w_0 w_1 \ldots w_{k-1}$

$$\sigma(w) = \lambda(w_0)\lambda(\{w_0, w_1\})\lambda(w_1) \ldots \lambda(w_{k-1})\lambda(\{w_{k-1}, w_0\}).$$

The set of *cyclic patterns* $\mathcal{C}(G)$ of a graph $G$ is defined by:

$$\mathcal{C}(G) = \{\pi(C) : C \in S(G)\}$$

where $S(G)$ denote the set of cycles of $G$. More information can be added to the kernel considering the set of the bridges of the graph. The set of tree patterns $T(G)$ assigned to graph $G$ is defined by:

$$T(G) = \{\pi(T) : T \text{ is a connected component of } B(G)\}$$

where $B(G)$ is the set of the bridges of the graph $G$. Assuming that $C(G)$ and $T(G)$ are disjoint, the cyclic pattern kernel is defined as follows:

**Definition 5.** *Let $G$ be a graph, the set $F(G)$ is defined as $C(G) \cap T(G)$. The cyclic pattern kernels between two graphs $G_1$ and $G_2$ is given by:*

$$K_{CP}(G_1, G_2) = |\mathcal{C}(G_1) \cap \mathcal{C}(G_2)| + |\mathcal{T}(G_1) \cap \mathcal{T}(G_2)|.$$

CPKs can be computed by the algorithm [13], whose input and output are respectively a graph $G$ and the set of cycles $S$, described below:

```
1. S=B= $\oslash$
2. Compute the set $\Gamma$ formed by the biconnected components of $G$
3. for $G' \in \Gamma$ do
4.     if $G'$ has one edge $e$ then $B = B \cup e$ else $S = S \cup C(G')$
5. $S = S \cup \{\pi(t) : t$ is a connected component of $B.\}$
6. return $S$
```

The biconnected components can be effectively computed by a depth-first search algorithm in a linear time [15]. Nevertheless, the computation of CPKs is *NP-hard* [13] mining, in this way, their real utility. To overcome the problem of intractability, Horvath et al. [13] proposed a restriction of CPKs. They observed that the simple cycles of a graph $G$ can be computed with polynomial complexity [16]. Besides, for a given graph $G$ and a $k \geq 0$, it can decide [16] if the number of cycles in $G$ is bounded by $k$. Using these facts, Horvath et al. [13] proposed to restrict the kernel to the case when the number of simple cycles is bounded by a constant for every graph in the database. This assumption is quite reasonable in real-world graph databases. They proposed a variant of the previous algorithm [13] that computes the cyclic pattern kernels only for graphs that have at most $k$ simple cycles. The variant, that takes as input a graph $G$ with $n$ vertices and $m$ edges and an integer $k \in \mathbb{N}$ and produces as output the set of cycles $S$, has the following steps:

```
1. $K = 0$
2. $S = B = \oslash$
3. Compute the set $\Gamma$ formed by the biconnected components of $G$
4. for $G' \in \Gamma$ do
5.     if $G'$ has more than one edge $e$ then
6.         $X$ =READ-TARJAN($G', k - K + 1$)
```

```
            if $|X|$ is equal to $k - K + 1$ then return $\oslash$
            else
              $S = S \cup \{p(C) : C \in X\}$
              $K = K + |X|$
        else $B = B \cup e$
  $S = S \cup \{\pi(t) : t$ is a connected component of $B.\}$
  return S
```

where READ-TARJAN is a procedure that implements the Read-Tarjan's algorithm that allows to decide if the number of cycles in a graph $G$ is bounded by $k$. SVMs with cyclic pattern kernels have been used effectively [13] on NCI-HIV database, a popular benchmark for the classification of chemical compounds.

## 7 The Kernel Approach

In this section we summarize the main peculiarities of Kernel Methods and we discuss the possible line of research. The main advantage of the kernel approach is its flexibility. The kernel methods approach for graphs results in designing an appropriate Mercer kernels on them. Computed the kernel, it can be used, without requiring further adjustments, in usual Kernel Methods for vectorial data. This allows to use the designed kernel either in supervised (e.g. SVMs, Gaussian Processes [7]) or in unsupervised Kernel Methods (e.g. *Kernel Clustering Methods* [17]). Although Kernel Methods for Graphs have been only investigated in their supervised form, Kernel Clustering Methods for graphs offer several potential advantages with respect to the unsupervised neural learning algorithms for graphs such as the SOM for structured data [2]. SOM can only produce piecewise linear separations among clusters. On the contrary, Kernel Clustering methods produce nonlinear separations allowing to tackle more highly nonlinear data sets. Therefore a challenging goal for Kernel Methods is represented by developing and validating Kernel Clustering Methods for graphs. Kernel Clustering Methods are not the unique algorithms able to produce nonlinear separations among clusters. Nonlinear separations can also be obtained by *Spectral Clustering Methods* [17]. The development of spectral clustering methods for graphs and, in general, for structured data remains an open problem.

## 8 Conclusion

The development of effective learning algorithms for structured data, i.e. data that cannot be represented by vectors, is one of the most challenging topic for machine learning. Recently Kernel Methods have tackled the structured data getting new force on the research about structured data. In this paper we have focused our attention on Kernel Methods that cope with data that can be represented by graphs. The aim of this paper has been to provide a comprehensive approach to discuss about the main Kernel Methods for graphs. One of the main challenging goals of Kernel Methods of graphs consists in extending the validation of these techniques, widely applied in bioinformatics and chemioinformatics, to other applicative areas such as computer vision applications.

## References

1. Frasconi, P., Gori, M., Sperduti, A.: A general framework for adaptive processing of data sequences. IEEE transactions on Neural Networks 9(5), 768–786 (1997)
2. Hagenbuchner, M., Sperduti, A., Tsoi, A.: A self-organizing map for adaptive processing of structured data. IEEE transactions on Neural Networks 14(23), 491–505 (2003)
3. Shawe-Taylor, J., Cristianini, N.: Kernels Methods for Pattern Analysis. Cambridge University Press, Cambridge (2004)
4. Gärtner, T.: A survey of kernels for structured data. SIGKDD Explorations 5(1), 49–58 (2003)
5. Berg, C., Christensen, J., Ressel, P.: Harmonic analysis on semigroups. Springer, New York (1984)
6. Gärtner, T., Flach, P., Wrobel, S.: On graph kernels: Hardness results and efficient alternatives. In: Proceedings of 16th Annual Conference on Computational Learning Theory and 7th Kernel Workshop, pp. 129–143. IEEE Press, Los Alamitos (2003)
7. Rasmussen, C., Williams, C.: Gaussian Processes for Machine Learning. MIT Press, Cambridge (2006)
8. Gärtner, T., Driessens, K., Ramon, J.: Graph kernels and gaussian processes for relational reinforcement learning. In: Horváth, T., Yamamoto, A. (eds.) ILP 2003. LNCS (LNAI), vol. 2835, pp. 146–163. Springer, Heidelberg (2003)
9. Gärtner, T., Lloyd, J., Flach, P.: Kernels for structured data. In: Matwin, S., Sammut, C. (eds.) ILP 2002. LNCS (LNAI), vol. 2583, pp. 66–83. Springer, Heidelberg (2003)
10. Kashima, H., Tsuda, K., Inokuchi, A.: Marginalized kernels between labeled graphs. In: Proceedings of 10th International Conference on Machine Learning, pp. 321–328. IEEE Press, Los Alamitos (2003)
11. Tsuda, K., Kin, T., Asai, K.: Marginalized kernels for biological sequences. Bioinformatics 18, S268–S275 (2002)
12. Mahé, P., Ueda, N., Akutsu, T., Perret, J.L., Vert, J.P.: Extensions of marginalized graph kernels. In: Proceedings of 21st International Conference on Machine Learning (ICML 2004), pp. 552–559. IEEE Press, Los Alamitos (2004)
13. Horvath, T., Gärtner, T., Flach, P., Wrobel, S.: Cyclic pattern kernels for predictive graph mining. In: Proceedings of KDD 2004, pp. 158–167. ACM Press, New York (2004)
14. Zaki, M.: Efficiently mining frequent trees in a forest. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 71–80. ACM Press, New York (2002)
15. Tarjan, R.: Depth-first search and linear graphs algorithms. SIAM Journal on Computing 1(2), 146–160 (1972)
16. Read, R., Tarjan, R.: Bounds on backtrack algorithms for listing cycles, paths and spanning trees. Networks 5(3), 237–252 (1975)
17. Filippone, M., Camastra, F., Masulli, F., Rovetta, S.: A survey of kernel and spectral method for clustering. Pattern Recognition 41(1), 174–192 (2008)

# Event-Based Compression and Mining of Data Streams

Alfredo Cuzzocrea[1] and Sharma Chakravarthy[2]

[1] ICAR Inst. and University of Calabria, Italy
cuzzocrea@si.deis.unical.it
[2] Dept. of Computer Science & Engineering
The University of Texas at Arlington
sharma@cse.uta.edu

**Abstract.** An innovative event-based data stream compression and mining model is presented in this paper. The main novelty of our approach with respect to traditional data stream compression approaches relies on the semantics of the application in driving the compression process by identifying "interested" events occurring in the unbounded stream. This puts the basis for a novel class of intelligent applications over data streams where the knowledge on actual streams is integrated with and correlated to the knowledge related to expired events that are considered critical for the target application scenario.

## 1 Introduction

The problem of efficiently representing [4] and mining [21] data streams is of relevant interest for both the Database and Data Mining research communities. Basically, data stream query processing poses novel and previously-unrecognized research challenges that make traditional DBMS technology (e.g, RDBMS) inadequate to the goal of dealing with their *unbounded nature*. In fact, while information stored in relational databases is represented by means of highly-detailed tuples, and query processing algorithms are *multi-step* accordingly, data streams cannot be represented in great detail as the stream is, potentially, unbounded. Also, data stream query processing algorithms typically operate within the context of specialized bounded *time windows*, which collect sets of stream readings (e.g., the last $T$ readings, with $T > 0$), with the constraint of applying in *one-pass* only.

To illustrate with an example, consider a sensor network that monitors environmental parameters located in a given geographic area. Focus on the simple case in which the temperature is monitored. Here, we can observe that for most part of the monitored interval of time, temperature readings are near to the average values observed during all the time (depending on the season). Nevertheless, there could happen sporadic events in which temperature readings are far from the average values (depending on sporadic atmospheric events). It should be noted that analysis of old data is indeed relevant in this particular application

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 670–681, 2008.

context, just like the analysis of new data. Also, this analysis model offers interesting perspectives for *cross and correlation analysis tools*, which play a leading role when complex data streams are considered [17]. In other words, *knowledge discovery methodologies over data streams can obtain a critical "plus-value" from the amenity of combining intermediate and final mining results over most recent readings and past readings that refer to expired significant events.*

The above-described analysis/mining perspectives over data streams lead to the definition of the so-called *event-based data stream compression and mining paradigm*, which can be reasonably considered as an innovative contribution over the state-of-the-art. In fact, to the best of our knowledge this aspect has not been considered by active literature previously.

The *synergistic* integration between data stream and event processing has been firstly highlighted in [28]. Basically, this proposal studies how the two research fields can be efficiently made synergic into powerful computation models for complex time-depending systems (like data stream processing systems). The results of this research effort have been synthesized in `MavEStream`, a four stage integration model for data streams and events.

Following this main intuition, in this paper we introduce the novel event-based data stream compression and mining paradigm discussed above, and we propose a formal model, called `ECM-DS` (`E`*vent-based* `C`*ompression and* `M`*ining of* `D`*ata* `S`*treams*), which allows us to meaningfully realize the proposed paradigm.

## 2 Synergistic Integration of Stream and Event Processing

*Event processing* [39, 15, 23, 22, 7, 10, 18, 16] and lately data stream processing [5,1,31,28,33,4,8,12,17,24,25,26,32,36] have evolved independently based on situation monitoring application needs. Several event specification languages [23,22,11,38,2,3] for specifying *composite events* have been proposed and triggers have been successfully incorporated into relational databases. Different computation models [22, 18, 7, 10, 16] for processing events, such as *Petri Nets* [22], *Extended Automata* [23], and *Event Graphs* [7,10,18] have been proposed and implemented. Various event *consumption (or detection) modes* [7,22,10,11], also called parameter contexts, have been explored. Similarly, data stream processing has received a lot of attention lately, and a number of issues from architecture [1,31,28,35] to *Quality-Of-Service* (QoS) [41,13,29,6,9] have been considered. Although both of these topics seem different on the face of it, one can see that there are more similarities than differences between them. Not surprisingly, the computation model used for data stream processing is not very dissimilar from some of the event processing models (e.g., event graph), but used with a different emphasis.

As many of the stream applications are based on sensor data, they invariably give rise to events on which some actions need to be taken. In other words, many stream applications seem to not only need computations on streams, but also these computations generate interesting events (e.g., car accident detection and notification, network congestion control, network fault management, intrusion detection), and several such events may have to be composed, detected, and monitored for taking appropriate actions. Currently, to the best of our knowledge, none of the work addresses the specification and computation of the above two threads of work. Our premise for this paper is that although each one is useful in its own right, their combined expressiveness and computation are critical for many applications of stream processing. Hence, there is a need for synthesizing the two into a more expressive and more powerful model that combines the strengths of each one.

Clearly, *it is desirable and natural to combine the strengths of both models into an integrated model with a general framework* and a set of comprehensive techniques of stream processing model plus the event computation model (i.e., computation at tuple level, consumption modes, and so on) and sophisticated rule processing capabilities. This integrated model will be much stronger and can serve a larger class of applications than what are currently supported by both the models individually. Inspired by these considerations, [28] proposes `MavEStream`, a four stage integration model for data streams and events, as mentioned in Sect. 1. `MavEStream` is a general-purpose model for event and data stream processing which can be easily customizable to different and even-heterogeneous application scenarios thanks to the amenity of defining the so-called semantic windows. With regards to the goals of `ECM-DS`, here the emphasis is on meaningfully introducing semantic windows able to capture the "degree of interestness" of events generated by data streams, and defining the corresponding SQL statements accordingly. This to gain advantages during both the compression and mining phases.

## 3 The Event-Based Compression Model for Data Streams

The event-based data stream compression and mining paradigm encompasses two models, namely the event processing model and the event-based compression model, according to the guidelines provided in Sect. 1. Sect. 3.1 focuses on the event processing model, whereas Sect. 3.2 deals with the event-based compression model.

### 3.1 Event Processing Model

*Event-Condition-Action* (ECA) rules are used to process event sequences and to make the underlying system active for applications such as situation monitoring, access control, and change detection. They consist of three components and they are (*i*) *Event*: occurrence of interest such as data-manipulation-events, clock-events, and external-notification-events; (*ii*) *Condition*: can be a simple or a complex query; (*iii*) *Action*: specifies the operations that are to be performed when an event occurs and the corresponding condition evaluates to true. ECA rules can be defined either at application level or system level. A number of event processing systems using ECA rules have been proposed and implemented in the literature: *ACOOD* [18], *ADAM* [15], *Alert* [39], *Ariel* [27], *COMPOSE* [23], *Hipac* [14], *ODE* [23], *REACH* [7], *Rock & Roll* [16], *SAMOS* [22], *Sentinel*

[10,11], *SEQ* [40], *UBILAB* [22], and [34]. A comprehensive introduction and description about most of these systems can be found in [42,37].

Several approaches have been proposed for the detection of composite events in the literature, among which we recall: *Event Detection Graphs* (EDGs) [18,7, 11], *Extended Finite State Automaton* [30], *Colored Petri Nets* [22], and *Event Algebra* [22]. EDGs have been shown to be based on operators rather than instances and hence are efficient as compared to other approaches based on the computation and storage requirements for detecting events.

In this paper, we draw upon EDGs for our integrated model as it corresponds to operator trees and has similarities with respect to query processing whereas the other representations do not share these characteristics with query processing. We also use the masking capability introduced in ODE [23] to filter events on arbitrary conditions.

It should be noted that EDGs, being based on reasoning models over operators and query processing, are particularly suitable to support our event-based data stream compression and mining paradigm, as this allows us to avoid excessive computational overheads that would be instead introduced if an instance-based model was employed. Also, this particular amenity successfully supports the *intrinsic* scalable nature of complex data stream processing applications and systems.

### 3.2 Event-Based Compression Model

Due to its particular data organization focused to capture the representation of two-dimensional aggregate information on time-evolving data streams, $MRDS$ is particularly suitable to support the non-linear compression paradigm, which represents a meaningfully extension of the original linear compression paradigm [12] that views “old” aggregate readings as less interesting than “new” aggregate readings.

Under the non-linear compression scheme, the $MRDS$ is compressed in dependence on a given degree of approximation $\delta$, which is obtained from the above-introduced event processing layer that elaborates semantic windows in order to adequately capture the degree of interestness of events and, then, produces in output this parameter. The way of defining and handling the relationship between interesting events and $\delta$ strongly depends on the particular data stream processing application scenario, and hence it must be determined and characterized accordingly during the start-up phase by the system administrator on the basis of his/her knowledge about the specific application domain. For this reason, $\delta$ must be considered as a free parameter of the event-based data stream compression model we propose, and must be contextualized to the particular instance. Therefore, in the rest of this Section we will treat $\delta$ as an input parameter, and we discuss the non-linear compression process accordingly.

Interesting events originated by data streams are continuously collected and recorded. Due to this task, the whole temporal dimension of the $MRDS$ is *annotated* by means of tuples of the following kind: $\langle E_k, t_{E_k,start}, t_{E_k,end}, \delta_k \rangle$, such that $(i)$ $E_k$ is the (interesting) event, $(ii)$ $t_{E_k,start}$ is the starting timestamp in which the event $E_k$ occurs, (*iii*) $t_{E_k,end}$ is the ending timestamp in which the event $E_k$ expires, (*iv*) $\delta_k$ is the degree of approximation required for the compression of aggregate values on data streams related to the event $E_k$. $\delta_k$ can be reasonably expressed as a percentage value. $\delta_k = 0\%$ means that the portion of aggregate values whose time interval is contained within the range $[t_{E_k,start} : t_{E_k,end}]$ must be maintained uncompressed, as $E_k$ is an event of particular relevance. $\delta_k = 100\%$ means that the portion of aggregate values whose time interval is contained within the range $[t_{E_k,start} : t_{E_k,end}]$ is not critical for the analysis goals of the specific knowledge discovery process over data streams considered, and hence this portion can be completely removed by saving the aggregate values of root nodes of the involved $QTW$ solely. Any other intermediate value of $\delta_k$ originates a *partial* compression of aggregate values whose time interval is contained within the range $[t_{E_k,start} : t_{E_k,end}]$. This overall determines an event-based compression of the $MRDS$, i.e. a compression that is driven by the degree of interestness and the relevance of events.

**Non-Linear Compression of** $QTW$**.** Before showing how the non-linear compression process works on the whole $MRDS$, we focus the attention on how a singleton $QTW$ is compressed, which is a basic task of the latter process. The tree-based nature of $QTW$ combined with the multi-resolution nature of OLAP queries offer a meaningful and intuitive way of modeling a partial compression of the $QTW$ in dependence on $\delta$. First, note that, as said above, for an uncompressed $QTW$, $\delta = 0\%$ whereas for a full-compressed $QTW$, $\delta = 100\%$. In the first case (i.e., $\delta = 0\%$), the $QTW$ maintains all the nodes and it is able to efficiently answer OLAP queries with the higher degree of approximation supported by the current representation of the $MRDS$. In the latter case (i.e., $\delta = 100\%$), the $QTW$ is reduced to its root node solely, and the evaluation of OLAP queries introduces high approximation given by applying classical linear interpolation techniques to the *unique* $\Delta S_{QTW} \times \Delta T_{QTW}$ two-dimensional range $\langle \mathcal{R}_{S,QTW}, \mathcal{R}_{T,QTW} \rangle$ associated to the root of the $QTW$.

Intuitively enough, from the model above it follows that a partial compression of the $QTW$ produces a partially-compressed $QTW$, denoted by $\widetilde{QTW}_p$, having a number of nodes between the two opposite situations represented by a full-compressed $QTW$, denoted by $\widetilde{QTW}_f$, with number of nodes $N_{\widetilde{QTW}_f} = 1$, and an uncompressed $QTW$, denoted by $\widetilde{QTW}_u$, with number of nodes $N_{\widetilde{QTW}_u} = \sum_{k=0}^{P_{\widetilde{QTW}_u}} 4^k$ (i.e., $1 < N_{\widetilde{QTW}_p} < \sum_{k=0}^{P_{\widetilde{QTW}_u}} 4^k$).

Another point of interest with respect to approximate answers to range-SUM queries on summarized data streams embedded into $QTW$ is noticing that *the accuracy of approximate answers is mostly due to the contribution of leaf nodes of the* $QTW$. This claim is obviously valid for those queries that overlap the two-dimensional ranges associated to internal nodes of the $QTW$, which are indeed the most frequent ones. Based on this main intuition, our partial $QTW$ compression task adopts the strategy of *pruning a number of leaf nodes, starting from the oldest ones, until the desired degree of approximation* $\delta$ *is satisfied.*

In order to dynamically detect the condition above, we introduce an *error metrics* over a "typical" synthetic query-workload against the current representation of the target $QTW$ to be compressed. Given the $QTW$ to be compressed, we introduce a query-workload $QWL$ against the $QTW$ composed by all those synthetic range-SUM queries $Q_s$ having area equal to $\Delta S_{Q_s} \times \Delta T_{Q_s}$, where $\Delta S_{Q_s}$ and $\Delta T_{Q_s}$ are input customizable parameters such that the following constraints are satisfied: $(i)$ for each synthetic query $Q_s$ in $QWL$, the two-dimensional range associated to $Q_s$ is contained by the two-dimensional range associated to the $QTW$ (i.e., $\langle \mathcal{R}_{S,Q_s}, \mathcal{R}_{T,Q_s} \rangle \subseteq \langle \mathcal{R}_{S,QTW}, \mathcal{R}_{T,QTW} \rangle$); $(ii)$ synthetic queries in $QWL$ overlap all the two-dimensional ranges associated to internal nodes of the $QTW$. Our error metrics is given by the *Average Relative Error* $\varepsilon_{r,QWL}$ due to evaluating synthetic queries in $QWL$ against the $QTW$, defined as follows: $\varepsilon_{r,QWL} = \frac{1}{|QWL|} \times \sum_{k=0}^{|QWL|-1} E_r(Q_{s,k})$, such that $E(Q_{s,k})$ is the relative approximation error due to evaluating the synthetic query $Q_{s,k}$ against the $QTW$, defined as follows: $E_r(Q_{s,k}) = \frac{|A(Q_{s,k}) - \widetilde{A}(Q_{s,k})|}{A(Q_{s,k})}$, where $A(Q_{s,k})$ $(\neq 0)$ is the exact answer to $Q_{s,k}$, i.e. the answer to $Q_{s,k}$ evaluated against the uncompressed data streams related to the $QTW$, and $\widetilde{A}(Q_{s,k})$ is the approximate answer to $Q_{s,k}$, i.e. the answer to $Q_{s,k}$ evaluated against the $QTW$.

Due to massive size of data streams, the above-illustrated algorithm could still introduce excessive computational overheads, so that some optimizations are necessary. Among all complexity aspects, data access cost is the most relevant one. Also, it should be noted that the input degree of approximation $\delta$ is indeed a *lower-bound*, meaning that the partially-compressed $QTW$, $\widetilde{QTW}_p$, must retain *at least* a degree of approximation equal to $\delta$. The latter two evidences suggest us *to prune sets of contiguous $QTW$ leaf nodes at time rather than one $QTW$ leaf node at time*, and check the approximation condition *after* that each set of contiguous $QTW$ leaf nodes has been removed from the $QTW$. The order in which the set of contiguous leaf nodes of the *actual* $QTW$ is selected is the temporal one, according to the motivations on the accuracy of approximate answers given above. The number of contiguous leaf nodes removed at each iteration, denoted by $\rho$, is a customizable input parameter of the partial compression task. Overall, this data access strategy involves in a lower spatio-temporal complexity during the partial compression task over the target $QTW$. Algorithm `partialCompress` (see Fig. 1) implements the partial compression task introduced above.

Finally, algorithm `fullCompress` (see Fig. 2) implements the full compression task introduced above. It is very simple and does not deserve further details.

**Non-Linear Compression of the *MRDS*.** The non-linear compression of the $MRDS$ exploits the non-linear compression of the $QTW$ as a baseline operation. Recall that the event process layer originates an annotation of the whole temporal dimension of the $MRDS$ by means of tuples of kind: $\langle E_k, t_{E_k,start}, t_{E_k,end}, \delta_k \rangle$. For the sake of simplicity, in the following analysis assume that each temporal range $[t_{E_k,start} : t_{E_k,end}]$ of tuples contains the temporal ranges of an *integer* number of $QTW$ of the $MRDS$. Extending the following analysis to the case in which each

```
Input: The $QTW$ to be compressed; the degree of approximation $\delta$; the width of the range of synthetic queries on the stream source dimension, $\Delta S_{Q_s}$; the width of the range of synthetic queries on the temporal dimension, $\Delta T_{Q_s}$; the number of contiguous leaf nodes to be removed at each iteration, $\rho$.
Output: Void.
Method: Perform the following steps:
$QWL \leftarrow computeQWL(QTW, \Delta S_{Q_s}, \Delta T_{Q_s})$;
$\widetilde{QTW}_p \leftarrow QTW$;
while($checkAppxDegree(\widetilde{QTW}_p, QWL, \delta)$ == FALSE && $\widetilde{QTW}_p.depth() \geq 1$){
  $\mathcal{L} \leftarrow nextContiguousLeafNodes(\widetilde{QTW}_p, \rho)$;
  $\widetilde{QTW}_p.remove(\mathcal{L})$;
}
$QTW \leftarrow \widetilde{QTW}_p$;
return;
```

**Fig. 1.** Algorithm `partialCompress`

```
Input: The $QTW$ to be compressed.
Output: Void.
Method: Perform the following steps:
$\widetilde{QTW}_f \leftarrow QTW.root()$;
$QTW \leftarrow \widetilde{QTW}_f$;
return;
```

**Fig. 2.** Algorithm `fullCompress`

temporal range $[t_{E_k,start} : t_{E_k,end}]$ contains the temporal ranges of a non-integer number of $QTW$ is straightforward.

Given the amount of storage space $B'$ to be released in order to represent "new" arrivals in the $MRDS$, the non-linear compression of the $MRDS$ works as follows. First, event annotation tuples are sorted by *descendent values* of $\delta_k$. This approach is due to observing that, given a tuple $\langle E_k, t_{E_k,start}, t_{E_k,end}, \delta_k \rangle$, a higher value of $\delta_k$ means that a higher degree of approximation is required for the aggregate values on data streams related to the temporal range $[t_{E_k,start} : t_{E_k,end}]$ (in other words, the event $E_k$ is relevant for the target application context). On the contrary, a lower value of $\delta_k$ means that a lower degree of approximation is required for the aggregate values on data streams related to the temporal range $[t_{E_k,start} : t_{E_k,end}]$ (in other words, the event $E_k$ is not critical for the target application context). Also, note that, in the first case, a lower amount of storage space can be released from the involved $MRDS$ portion (i.e., a set of $QTW$), whereas, in the second case, a higher amount of storage space can be released from the involved $MRDS$ portions.

**Input:** The $MRDS$ to be compressed; the amount of storage space to be release, $B'$.
**Output:** Void.
**Method:** Perform the following steps:

```
$EventTupleSet \leftarrow MRDS.eventTupleSet()$;
$EventTupleSet.sortByAppxDegree()$;
$\widetilde{MRDS} \leftarrow MRDS$;
for($k = 0 \ldots EventTupleSet.size()$ && $B' > 0$){
  $EventTuple \leftarrow EventTupleSet.eventTuple(k)$;
  $t_{E_k,start} \leftarrow EventTuple.tStart()$;
  $t_{E_k,end} \leftarrow EventTuple.tEnd()$;
  $\delta_k \leftarrow EventTuple.delta()$;
  $QTWSet \leftarrow \widetilde{MRDS}.QTWSet(t_{E_k,start}, t_{E_k,end})$;
  while($h = 0 \ldots QTWSet.size()$ && $B' > 0$){
    $QTW \leftarrow QTWSet.QTW(h)$;
    if($\delta_k \neq 0$){
     if($\delta_k \doteq 100$){
       $QTW.fullCompress()$;
     else
       $QTW.partialCompress(\delta_k)$;
     }
     $B' \leftarrow B' - computeReleasedSpace(\widetilde{MRDS}, MRDS)$;
    }
  }
}
$MRDS \leftarrow \widetilde{MRDS}$;
return;
```

**Fig. 3.** Algorithm `compress`

The storage space $B'$ to be released is indeed a *lower-bound*, meaning that during the $MRDS$ compression process we could finally release a total amount $B'' > B'$, since additional space needed to the representation of structural information of the $MRDS$ could be removed while removing $QTW$ nodes (with the goal of releasing $B'$). However, the additional space $B''' = B'' - B'$ can be invested then to represent new other arrivals. Another specific feature of the $MRDS$ compression process relies in the relationship between the latter process with the basic task represented by the compression of a singleton $QTW$. Since stream readings are *buffered* before to be stored within the $MRDS$ (to performance purposes), at each iteration of the overall compression we need to compress several $QTW$, according to the "relevance" of associated events, until the required storage space $B'$ is completely released.

This evidence suggests us to (fully) compress those $QTW$ requiring a degree of approximation equal to 100% first, and (partially) compress those $QTW$ requiring a degree of approximation between 0% and 100% then. Obviously, $QTW$ requiring a degree of approximation equal to 0% are maintained uncompressed. Overall, this strategy allows us to release the required amount of storage space

$B'$, and, at the same time, ensure that the constraints imposed by the event-based data stream compression model can be satisfied. In fact, it should be noted that full $QTW$ compressions allow us to release a higher amount of storage space and, as a consequence, the amount of storage space to be released during partial $QTW$ compressions can be lowered, thus involving a better approximation in those $QTW$ related to interesting events.

Finally, algorithm `compress` (see Fig. 3) implements the non-linear $MRDS$ compression process illustrated above.

## 4 A Reference Architecture for Event-Based Compression and Mining of Data Streams

The reference architecture implementing the proposed `ECM-DS` model is shown in Fig. 4 and it consists of four stages (as a straightforward extension of the `MavEStream` architecture [28] focused to the compression of data streams): 1) CQ processing stage used for computing CQs over data streams; 2) event generation stage in which interesting events are generated as a result of continuous query processing; 3) event processing stage that is used for detecting events with/without masks; 4) rule processing stage that is used to check conditions, and to trigger predefined *compression* actions once events are detected.

The seamless nature of our integrated model is due the compatibility of the chosen event processing model (i.e., an event detection graph) with the structure used for stream processing. Based on our analysis, synthesizing both the processing models with respect to the goals of `ECM-DS` requires the following issues to be addressed: 1) handling highly bursty event streams (generated by the CQ processing stage) in event processing; 2) processing of events streams based on attributes and not solely on timestamp; 3) specification of events/event expressions, rules and CQs.

In order to efficiently support `ECM-DS`, we have enhanced both the models to address the above mentioned issues. For the event processing model, we performed the following improvements: 1) we have enhanced the event operators by introducing input queue(s) for each operator, which makes it possible to handle highly bursty outputs from CQ processing stage and take advantages of the techniques (i.e., scheduling strategies,load shedding) developed for stream processing model; 2) we have enhanced the event expressions in such a way that primitive events can process event streams based on event attributes, and not only on timestamp; 3) we have also enhanced the event consumption modes to support more meaningful windows; 4) we have extended SQL allowing user to specify events/event expressions, rules and CQs together.

For the stream processing model: 1) output of CQs can be fed as inputs to the primitive events in the event processing stage – we have named continuous queries, so that in the event processing stage the outputs of CQs can be used for detecting events; 2) we have introduced stream modifiers that can detect complex changes between tuples in a stream; 3) we have also introduced the

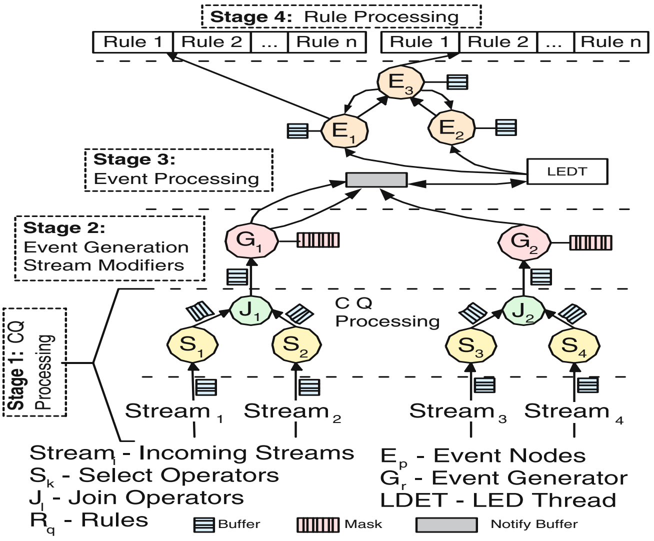


**Fig. 4.** ECM-DS reference architecture

semantic window to enhance the expressiveness and computation efficiency of CQs, and to allow creation of more meaningful windows.

## 5 Conclusions and Future Work

ECM-DS not only supports a larger class of data stream processing and mining applications, but also provides more accurate and efficient ways for processing CQs and event expressions. All the enhancements proposed in this paper deal seamlessly with current stream processing techniques and can be easily integrated into any current data stream management systems. Finally, a prototype of the proposed integrated model is underway.

In terms of future work, we are investigating how ECM-DS can be integrated with novel *Grid architectures* [20, 19] in order to embed into the model high performance and high reliability features that will allow us to efficiently process high-rate, high-dimensional, massive data streams.

## References

1. Abadi, D., et al.: Aurora: A New Model and Architecture for Data Stream Management. VLDB Journal 12(2) (August 2003)
2. Adaikkalavan, R., Chakravarthy, S.: SnoopIB: Interval-Based Event Specification and Detection for Active Databases. In: Proceedings, East-European Conference on Advances in Databases and Information Systems (September 2003)

3. Adaikkalavan, R., Chakravarthy, S.: Formalization and Detection of Events Over a Sliding Window in Active Databases Using Interval-Based Semantics. In: Proceedings, East-European Conference on Advances in Databases and Information Systems (September 2004)
4. Babcock, B., Babu, S., Datar, M., Motwani, R., Widom, J.: Models and issues in data stream systems. In: ACM PODS (2002)
5. Babu, S., Widom, J.: Continuous Queries over Data Streams. In: ACM SIGMOD RECORD (September 2001)
6. Brian, B., et al.: Chain: Operator Scheduling for Memory Minimization in Stream Systems. In: Proceedings, International Conference on Management of Data (SIGMOD) (2003)
7. Buchmann, A.P., et al.: Rules in an Open System: The REACH Rule System. Rules in Database Systems (1993)
8. Cai, Y., Clutterx, D., Papex, G., Han, J., Welgex, M., Auvilx, L.: Maids: Mining alarming incidents from data streams. In: ACM SIGMOD (2004)
9. Carney, D., et al.: Operator Scheduling in a Data Stream Manager. In: Proceedings, International Conference on Very Large Data Bases (September 2003)
10. Chakravarthy, S., et al.: Design of Sentinel: An Object-Oriented DBMS with Event-Based Rules. Information and Software Technology 36(9), 559–568 (1994)
11. Chakravarthy, S., Mishra, D.: Snoop: An Expressive Event Specification Language for Active Databases. Data and Knowledge Engineering 14(10), 1–26 (1994)
12. Cuzzocrea, A., Furfaro, F., Masciari, E., Saccà, D., Sirangelo, C.: Approximate Query Answering on Sensor Network Data Streams. In: Stefanidis, A., Nittel, S. (eds.) GeoSensor Networks (2004)
13. Das, A., Gehrke, J., Riedewald, M.: Approximate Join Processing over Data Streams. In: Proceedings, International Conference on Management of Data (SIGMOD) (2003)
14. Dayal, U., et al.: The HiPAC Project: Combining Active Databases and Timing Constraints. SIGMOD Record 17(1), 51–70 (1988)
15. Diaz, O., Paton, N., Gray, P.: Rule Management in Object-Oriented Databases: A Unified Approach. In: Proceedings, International Conference on Very Large Data Bases (September 1991)
16. Dinn, A., Williams, M.H., Paton, N.W.: ROCK & ROLL: A Deductive Object-Oriented Database with Active and Spatial Extensions. In: Proceedings, International Conference on Data Engineering (1997)
17. Dobra, A., Gehrke, J., Garofalakis, M., Rastogi, R.: Processing complex aggregate queries over data streams. In: ACM SIGMOD (2002)
18. Engstrom, H., Berndtsson, M., Lings, B.: Acood essentials. Technical report, University of Skovde (1997)
19. Foster, I., Kesselman, C., Nick, J., Tuecke, S.: Grid Services for Distributed System Integration. IEEE Computer 35(6) (2002)
20. Foster, I., Kesselman, C., Tuecke, S.: The Anatomy of the Grid: Enabling Scalable Virtual Organizations. International Journal of High Performance Computing Applications 15(3) (2001)
21. Gaber, M., Zaslavsky, A., Krishnaswamy, S.: Mining Data Streams: A Review. ACM SIGMOD Record 34(2) (2005)
22. Gatziu, S., Dittrich, K.R.: Events in an Object-Oriented Database System. In: Proceedings of Rules in Database Systems (September 1993)
23. Gehani, N.H., Jagadish, H.V., Shmueli, O.: Composite Event Specification in Active Databases: Model & Implementation. In: Proceedings, International Conference on Very Large Data Bases, pp. 327–338 (1992)

24. Gehrke, J., Korn, F., Srivastava, D.: On computing correlated aggregates over continual data streams. In: ACM SIGMOD (2001)
25. Gilbert, A., Kotidis, Y., Muthukrishnan, S., Strauss, M.: One-Pass Wavelet Decompositions of Data Streams. IEEE Trans. on Knowledge and Data Engineering 15(3) (2003)
26. Guha, S., Koudas, N., Shim, K.: Data streams and histograms. In: ACM STOC (2001)
27. Hanson, E.N.: Active Rules in Database Systems, pp. 221–232. Springer, New York (1999)
28. Jiang, Q., Chakravarthy, S.: Data Stream Management System for MavHome. In: Proceedings, Annual ACM Symposium on Applied Computing (March 2004)
29. Jiang, Q., Chakravarthy, S.: Scheduling Strategies for Processing Continuous Queries over Streams. In: Proceedings, British National Conference on Databases (July 2004)
30. Lieuwen, D.L., Gehani, N.H., Arlein, R.: The Ode Active Database: Trigger Semantics and Implementation. In: Proceedings, International Conference on Data Engineering, March 1996, pp. 412–420 (1996)
31. Madden, S., Franklin, M.J.: Fjording the Stream: An Architecture for Queries over Streaming Sensor Data. In: Proceedings, International Conference on Data Engineering (2002)
32. Manku, G., Motwani, R.: Approximate frequency counts over data streams. In: VLDB (2002)
33. Mokbel, M.F., et al.: PLACE: A Query Processor for Handling Real-time Spatio-temporal Data Streams. In: Proceedings, International Conference on Very Large Data Bases
34. Motakis, I., Zaniolo, C.: Temporal Aggregation in Active Database Rules. In: Proceedings, International Conference on Management of Data (SIGMOD), pp. 440–451 (1997)
35. Motwani, R., et al.: Query Processing, Resource Management, and Approximation in a Data Stream Management System. In: Proceedings, Conference on Innovative Data Systems Research (January 2003)
36. Muthukrishnan, S.: Data streams: Algorithms and applications. In: ACM-SIAM SODA (2003)
37. Paton, N.W.: Active Rules in Database Systems. Springer, New York (1999)
38. Roncancio, C.: Toward Duration-Based, Constrained and Dynamic Event Types. In: Active, Real-Time, and Temporal Database Systems, pp. 176–193 (1997)
39. Schreier, U., et al.: Alert: An Architecture for Transforming a Passive DBMS into an Active DBMS. In: Proceedings, International Conference on Very Large Data Bases (1991)
40. Seshadri, P., Livny, M., Ramakrishnan, R.: The Design and Implementation of a Sequence Database System. In: Proceedings, International Conference on Very Large Data Bases, pp. 99–110 (1996)
41. Tatbul, N., et al.: Load Shedding in a Data Stream Manager. In: Proceedings, International Conference on Very Large Data Bases (September 2003)
42. Widom, J., Ceri, S.: Active Database Systems: Triggers and Rules. Morgan Kaufmann Publishers, San Francisco (1996)

# Context Enabled Semantic Granularity

Riccardo Albertoni[1], Elena Camossi[2,⋆], Monica De Martino[1],
Franca Giannini[1], and Marina Monti[1]

[1] IMATI, Consiglio Nazionale delle Ricerche, Via De Marini, 6 - Torre di Francia - 16149
Genova, Italy
{albertoni,demartino,giannini,monti}@ge.imati.cnr.it

[2] School of Computer Science and Informatics, University College Dublin, Belfield,
Dublin 4, Ireland
elena.camossi@ucd.ie

**Abstract.** In this paper we propose a powerful ontology driven method that eases the browsing of any repository of information resources described by an ontology: we provide a flexible semantic granularity method for the navigation of a repository according to different levels of abstraction, i.e. granularities. The granularity is explicitly parameterised according to the criteria induced by the context.

## 1 Introduction

*Semantic granularity* enables the browsing of information resources according to different levels of abstraction, i.e., granularities. Granularities have been already studied in the area of Information Systems, in particular for the spatio-temporal domain [1]. Some attempts to define semantic granularities have been made with respect to terminologies. However, in both cases, granularities are static and embedded in the data model or in the database schema. By contrast, semantic granularity [2] extracts dynamically the structure, namely the *granularity lattice*, which enables to organize the repository at different levels of abstraction.

Moreover, to fill the gap between Cognitive and Information spaces [2], it is mandatory to take into account the influence of the *context*. Thus, in this paper we propose a context dependent semantic granularity method. It originates from the research results presented in [3], where the *application context* has been formalized in order to parameterize the semantic similarity among ontology instances. The application context models the importance of ontology entities (i.e., classes, attributes and relations) that concur in the granularity assessment as well as the different operations used to analyse them. Herein we adapt this formalization to parameterize the semantic granularity we have proposed in [2]. The resulting instrument is a powerful ontology driven method that eases the browsing of a repository of information resources.

The advances of this work with respect to our previous results [3,2] are: (i) the layered framework becomes a potential common frame for context dependent ontology driven methods: we demonstrate it is suitable for both the semantic similarity and the

⋆ The work of Elena Camossi is supported by the Irish Research Council for Science, Engineering and Technology.

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 682–688, 2008.

granularity; (ii) we propose an extension of the application context formalism for the granularity: we define new operations and functions to be adopted for the analysis of ontology entities; (iii) we illustrate a more flexible evaluation of semantic granularity throughout its context dependent parameterization. Overall, the main benefit of this work is to enable a user-oriented browsing: the user may formulate, learn and modify the granularity criteria induced by the context.

The paper is organized as follows. In Section 2 we introduce the semantic granularity method discussing an illustrative example. In Section 3 we describe the context dependent parameterization of the method. Finally, Section 4 concludes the paper outlining future research directions.

## 2 Semantic Granularity

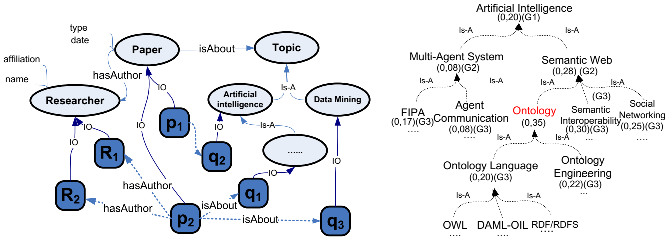
Semantic granularities are built with respect to an ontology $\mathcal{O}$ representing the information resources, which are described by a structured set of *qualities* $\mathcal{Q}$. Information resources are instances of a class $\mathcal{S}$. The set of qualities $\mathcal{Q}$ are represented by ontology classes organized in a hierarchy $\prec_Q$ induced by relations IS-A and Part-Whole. We suppose that a top hierarchy class $Q^T$ exists such that, for each quality $Q$, $Q \prec_Q Q^T$; moreover, each $Q \in \mathcal{Q}$ has at least one direct instance.

The user is expected to access the resources in the repository by using set of granules with increasing detail. Each granule belongs to a given granularity and corresponds to a quality $Q$ according to which the corresponding resources are grouped. Granularities are defined dynamically, according to both the data model, represented by the ontology schema, and the data, given by ontology instances.

The method follows a two-phase process. In the first phase, namely *quality filtering*, it evaluates each quality with respect to its capability of abstracting information resources. The evaluation of the *abstraction capability of a quality* $Q$ takes into account the attributes and the relations that characterize the resources in $\mathcal{S}$ as well as the attributes and the relations of their related instances. The quality filtering returns the qualities with a better value of abstraction capability which are promoted to be granules of some granularity.

Then, the *granularity building* phase distributes the granules among different granularities according to $\prec_Q$. It returns the set of granularities to employ for the repository navigation. Since not all the qualities in the hierarchy will be evaluated as good abstractors by the quality filtering phase, the browsing of the information resources according to semantic granularities will differ from the browsing driven by IS-A and Part-Whole.

*Example 1. Fig. 1 shows an example of application of semantic granularity onto a repository of scientific papers represented by the ontology schema in Fig. 1(a). We use instances of* Paper *as resources and of* Topic *as qualities in input for the semantic granularity. Fig. 1(b) is an excerpt of the topic taxonomy: the values in brackets are the results of the semantic granularity application. The first value is the abstraction capability of the topic resulting by the quality filtering phase: the lower is the value, the better the topic abstracts its subtopics in the hierarchy. Setting an abstraction threshold (for instance 0,31), the quality/topic* Ontology *is discarded. The second element*



(a) (b)

**Fig. 1.** (a) An ontology schema to organize information about scientific papers. (b) Topic taxonomy and semantic granularity results. For each quality, the values in brackets indicate its abstraction capability and the granularity (G1,G2, or G3) to which it is assigned.

*in the brackets represents the result of the granularity building phase: the granularities G1, G2, G3, which correspond to distinct levels of abstraction, are identified and the granules are associated with them. For example, increasing the level of detail,* Artificial Intelligence *belonging to G1 is converted in* Multi-Agent System *and* Semantic Web*, which belong to G2. Furthermore,* Semantic Web *is converted in* Ontology Language*,* Ontology Engineering*,* Semantic Interoperability*, and* Social Networking *belonging to G3.*

## 3 Context Dependent Parameterization of Semantic Granularity

### 3.1 The Ontology Model and the Layered Framework

The ontology model gives the expressiveness of the ontologies defined according to the framework. Herein, we adopt the ontology model equivalent to an ontology with data types and defined in [3]. In addiction to $\delta_a$, $\delta_r$, $\delta_c$ that retrieve the attributes, the relations and the concepts reachable by a given concept or relation, we defined the function $\delta_{r^{-1}}$: $C \cup R \rightarrow 2^R$, such that $\delta_{r^{-1}}(c)$=$\{r$: $R \mid \exists c' \in C, \sigma_R(r) = (c', c)\}$ denotes the set of relations that reach $c \in C$; and $\delta_{r^{-1}}(r)$=$\{r'$: $R \mid r' \neq r$, $\exists c \in C, \exists c' \in \delta_c(r), \sigma_R(r') = (c, c')\}$ is the set of relations which differ from $r$ and reaches the concepts reachable through the relation $r \in R$.

The framework is structured in terms of *data*, *ontology* and *context* layers plus the *domain knowledge* layer which spans all the others [4].

The *data layer* provides the *functions* onto the data type values (e.g., functions which filter the values of simple or complex data types, statistical and user defined functions).

The *ontology layer* provides the mechanism for processing semantic granularity by considering the way ontology's entities are related. It provides the implementation of the semantic granularity and of the *operations* (e.g., intersection, count) which may be recalled by the semantic granularity in a given application context.

The *context layer* provides the *application contexts*, i.e., the criteria for the computation of semantic granularity considering how ontology entities are used for specific purposes. Each application context specifies the attributes and the relations to consider likewise the operations and functions to apply on them.

### 3.2 Application Contexts

This section formalizes the application contexts used to parameterize the semantic granularity. It is an extension of the formalization illustrated in [3]. An application context is defined by an ontology engineer, according to specific application needs. Assuming the definition of *Sequence of elements* presented in [3] a *path of recursion* tracks the recursion during the assessment of the semantic granularity and represents the navigation path in the ontology to collect the information of interest. It is defined as follows.

**Definition 1 (Path of Recursion of length n).** *A* path of recursion $p$ *of length $n$ is a sequence of elements with length $n$ whose elements are classes in $C$ and relations in $R$ (i.e., $p \in S^n_{C \cup R}$), such that p starts from a class c and whose other elements are relations either starting from or ending in c or c′, where c′ is a class involved in some relation in p, that is* $p(1) \in C \wedge \forall j \in [2, n]\, p(j) \in R \wedge (p(j) \in \delta_r(p(j-1)) \vee p(j) \in \delta_{r^{-1}}(p(j-1)))$.

$P^n$ denotes the set of all paths of recursion with length $n$, whereas $P$ denotes the set of all paths of recursion $P = \bigcup_{n \in N} P^n$.

The *application context (AC)* function is defined inductively according to the length of the path of recursion. It yields the set of attributes and relations to consider and the operations to apply when computing the semantic granularities, e.g., sum, average, minimum, maximum, which could indirectly recall the functions in the data layer, and different forms of count operations: $Count$, which evaluates the cardinality of a set of instances; $WCount$, which evaluates a weighted count of instances according to the cardinality of related attributes or relations; $InvCount$, which evaluates the inverse cardinality of a set of instances, (i.e., a set with less instances has more importance than a set with greater cardinality). The application context is formally defined as follows.

**Definition 2 (Application Context AC).** *Given the set $P$ of paths of recursion, $L$ the set of operations provided by the ontology layer (i.e. Count, WCount and InvCount for the semantic granularity), G the set of datatype functions available in the data layer, the application context for the semantic granularity is defined by the partial function* $AC$: $P \rightarrow 2^{A \times (L \cup G)} \times 2^{R \times L}$.

Note that each application context $AC$ is characterized by the operators $AC_A$: $P \rightarrow 2^{A \times (L \cup G)}$ and $AC_R$: $P \rightarrow 2^{R \times L}$, which yield respectively the context $AC$ related to the attributes and to the relations.

*Example 2. Given the ontology schema in Fig. 1, two examples of application contexts $AC_1$ and $AC_2$ are defined. $AC_1$ corresponds to the hard coded context implicitly used in Example 1. It starts from the path of recursion $[Topic]$ and considers the instances of* Paper *associated with each* Topic *to calculate the capability of abstraction. It is formalised as follows:*

$$[Topic] \overset{AC_1}{\rightarrow} \{\{\phi\}, \{(isAbout^{-1}, Count)\}\}.$$

*$AC_2$ considers the date of publication, the number of authors, the type (i.e., journal, conference proceedings, or book) of papers. It is formalised as follows:*

$$[Topic] \overset{AC_2}{\rightarrow} \{\{\phi\}, \{(isAbout^{-1}, WCount)\}\}$$

$$[Topic.isAbout^{-1}] \overset{AC_2}{\rightarrow} \{\{(type, i(Paper, Book))(date, g(today))\} \\ \{(hasAuthor, InvCount)\}\} .$$

*$AC_2$ starts from the path of recursion $[Topic]$ and moves along to the inverse of relation* isAbout *to focus on the attributes and relations of* Paper. *The change of focus is tracked by the path of recursion $[Topic.isAbout^{-1}]$. $AC_2$, when applied to the new path of recursion, returns the attributes* type *and* date *and the relation* hasAuthor. Type *and* date *are respectively processed by the two functions $i$ and $g$ provided by the data layer: $i(Paper, Book)$ returns the inverse of the cardinality of the papers associated with a given topic that are published in a book or a journal, whereas $g(today)$ counts only the papers published in the last three years. Finally, the inverse cardinality of the relation* hasAuthor *is considered.*

### 3.3 Context Dependent Quality Filtering

As mentioned above, the *quality filtering* evaluates the abstraction capability of each quality $Q$, selecting those more representative for the repository that will become granules. We explicitly parameterize the capability of abstraction of a quality $Q$ to provide a semantic granularity according to an application context $AC$.

**Definition 3 (Abstraction capability of a quality $Q$ w.r.t. $AC$, $R_Q^{AC,p}$).** *Given an ontology $\mathcal{O}$, representing the class of information resources $\mathcal{S}$, described with respect to the set of qualities $\mathcal{Q}$; the quality $Q \in \mathcal{Q}$; the partial order $\prec_Q$ on $\mathcal{Q}$; $Q^T$ the most generic quality in $\mathcal{Q}$ according to $\prec_Q$; the starting path of recursion $p$ initialized as $p = [Q^T]$; $R_Q^{AC,p}$, is defined as follows:*

$$R_Q^{AC,p} = \frac{\sum_{x \in AC_A(p) \cup AC_R(p)} R_Q^{AC,p,x}}{|AC_A(p)| + |AC_R(p)|}.$$

$R_Q^{AC,p,x}$ is the abstraction capability of $Q$ according to the relation or attribute $x$ mentioned in the application context $AC$ for a path of recursion $p$. It is defined as follows:

$$R_Q^{AC,p,x} = \begin{cases} R_Q^{AC,p\circ x} & \text{if (x,}WCount\text{)} \in AC_R(p) \\ \frac{\max_{\{Q'|Q' \prec_Q Q\}} s_{AC,p,x}^{Q'*}}{\sum_{\{Q'|Q' \prec_Q Q\}} s_{AC,p,x}^{Q'*} + s_{AC,p,x}^{Q}} & \text{otherwise.} \end{cases}$$

In practice, for each relation $x$ in the context whose associated operation is $WCount$, the evaluation of $R_Q^{AC,p}$ is defined recursively considering the instances related to the quality $Q$ by the path of recursion $p \circ x$. That allows to assess the abstraction capability of $Q$ also considering the relations and attributes belonging to instances that are not directly related to the quality $Q$. Otherwise, when the context does not prescribe a recursive assessment, the abstraction capability presented in [2] is parameterized according to the context defining $s_{AC,p,x}^{Q}$ and $s_{AC,p,x}^{Q^*}$ as follows:

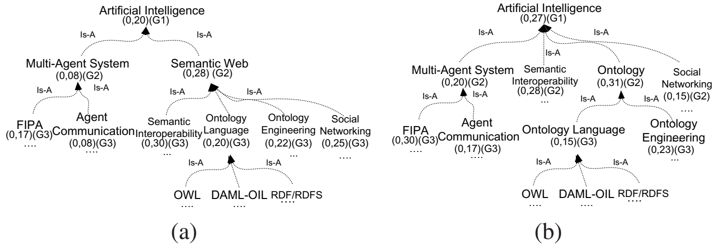
$$s_{AC,p,x}^{Q} = \sum_{q \in Q} \sum_{\iota \in I(q,p)} f_{AC}^{p,x}(\iota) \qquad\qquad s_{AC,p,x}^{Q^*} = \sum_{\{Q'|Q' \prec_Q Q\}} s_{AC,x}^{Q'}.$$

$f_{AC}^{p,x}(\iota)$ measures the *weight* of the instance $q$ of $Q$ according to the relation or attribute $x$ belonging to the set of instances $I(q,p)$, which are reachable through the recursion path $p$, and considering the operations indicated in $AC$. Thus, assuming: (i) $X$ a placeholder that works as a metasymbol that may be replaced by $R$ or $A$, whether $x$ is respectively a relation or an attribute; (ii) $i_A(\iota, a) = \{v \in V \mid (\iota, v) \in l_A(a), \exists y \in C \; s.t. \; \sigma_A(a) = (y, T) \wedge l_T(T) = 2^V\}$ the set of values assumed by the instance $\iota$ for attribute $a$; (iii) $i_R(\iota, r) = \{\iota' \in l_c(c') \mid \exists c \; \iota \in l_c(c) \; \exists \, c' \; s.t. \; \sigma_R(r) \in (c, c') \wedge (\iota, \iota') \in l_R(r)\}$ the set of instances related to the instance $\iota$ by relation $r$; (iiii) $g$ a function provided by the data layer, $w$ the metasymbol that works as placeholder for the function parameters that have been already fixed in the application contexts; $f^{p,x}_{AC}(\iota)$ is defined as follows:

$$f^{p,x}_{AC}(\iota) = \begin{cases} g(w) & \text{if}_{(x,g(w)) \in AC_A(p),\; v \in i_A(\iota,x)} \\ |i_X(\iota, x)| & \text{if } _{(x,\text{Count}) \in AC_X(p)} \\ \frac{1}{|i_X(\iota,x)|+1} & \text{if } _{(x,InvCount) \in AC_X(p)} \end{cases}$$

The following example gives the flavour of circumstances where different contexts arise.

*Example 3. We provide an example of semantic granularity application according to two application contexts. Let us considering a user who needs to browse a repository of scientific papers with two distinct purposes: (1) to get a first impression about the repository content and (2) to identify the hottest topics in the Computer Science research. These two aims correspond to two distinct contexts formalized respectively by the functions $AC_1$ and $AC_2$ given in the Example 2. We have extracted some real data from Faceted Dblp, a repository of Computer Science papers*[1]*, and organized them in the ontology of Fig. 1. The semantic granularity has been applied and a fragment of the result is illustrated in Fig. 2. Different granularities are obtained considering the contexts $AC_1$ and $AC_2$: comparing Fig. 2(a) and Fig. 2(b) granularities G2 and G3 contain different granules, i.e., topics. Indeed, the filtering phase results in different abstraction capabilities for the topics, according the two contexts. For example,* Semantic Web *disappears in Fig. 2(b), as it has been discarded by the filtering, whereas* Ontology *increases its importance: moving from $AC_1$ to $AC_2$,* Semantic Web *decreases its abstraction capability as $R_Q$ increases from 0.28 to 0.35, whereas* Ontology *increases its importance as $R_Q$ decreases from 0.35 to 0.31. Adopting the threshold 0.31, the granularity building phase returns the granularities depicted in Fig 2.*



(a) (b)

**Fig. 2.** Results of semantic granularity: according to contexts $AC_1$ (a) and $AC_2$ (b)

[1] Faced Dblp is available at http://dblp.l3s.de/.

## 4 Conclusions and Future Works

In this paper we have described a context dependent parameterization of semantic granularity to browse any kind of information source described with respect to a set of qualities represented in ontologies. Even if the work is at a preliminary stage, the intermediate results indicate the validity of the undertaken approach towards the definition of a powerful ontology driven method allowing a user-oriented formulation of the granularity criteria induced by the context.

A more rigorous test case is in progress. So far, the context as explicitly parameterization of ontology driven methods has been demonstrated to be essential both for the semantic similarity [3] and the granularity.

## References

1. Camossi, E., Bertolotto, M., Bertino, E.: A Multigranular Object-oriented Framework Supporting Spatio-temporal Granularity Conversions. International Journal of Geographical Information Science 20(5), 511–534 (2006)
2. Albertoni, R., Camossi, E., De Martino, M., Giannini, F., Monti, M.: Semantic Granularity for the Semantic Web. In: Meersman, R., Tari, Z., Herrero, P. (eds.) OTM 2006 Workshops. LNCS, vol. 4278, pp. 1863–1872. Springer, Heidelberg (2006)
3. Albertoni, R., De Martino, M.: Asymmetric and Context-Dependent Semantic Similarity among Ontology Instances. In: Spaccapietra, S. (ed.) Journal on Data Semantics X. LNCS, vol. 4900, pp. 1–30. Springer, Heidelberg (2008)
4. Ehrig, M., Haase, P., Hefke, M., Stojanovic, N.: Similarity for Ontologies - A Comprehensive Framework. In: Proc. of the 13th European Conf. on Information Systems, Information Systems in a Rapidly Changing Economy (2005)

# Modeling Spatial Relations between Places to Support Geographic Information Retrieval

Alia I. Abdelmoty[1] and Baher A. El-Geresy[2]

[1] Cardiff School of Computer Science,
Cardiff University, Wales, UK
[2] School of Computing,
University of Glamorgan, Wales, UK

**Abstract.** Geo-ontologies have a key role to play in the development of the geospatial-semantic web, with regard to facilitating the search for geographical information and resources. They normally hold large amounts of geographic information and undergo a continuous process of revision and update. This paper proposes a novel qualitative representation scheme to structure geo-ontology A-Boxes. The scheme captures two types of spatial relationships between geo-objects, namely, adjacency and orientation. It facilitates qualitative manipulation of the data and the efficient derivation of implicit information using spatial reasoning techniques.

## 1 Introduction

Almost eighty percent, as commonly quoted, of information on the web are geographically referenced. The simplicity and effectiveness of applications such as Google Earth led to a hype of activity to geo-referencing information on the web. A large number of documents stored and retrieved include references to geographic information, typically by means of place names and spatial relationships to place names. Both types of information play a key role in query formulation and information retrieval on the web.

Qualitative spatial relations are paramount in geographic information retrieval (GIR) systems. When expressing spatial relationships, people will use mostly qualitative vocabulary, such as "in", "north-of" and "near". The retrieval of this information requires a representation of places and their spatial extents. Approximate or relative information of spatial structure can be sufficient to retrieve the required qualitative relationships. Storage of every possible instance of relationships between places is not feasible and geometric computation carries an overhead and generally requires precise definition of spatial boundaries of these objects.

In this paper, we propose a model of Places and spatial relations that supports the explicit representation of spatial structure to facilitate the the realisation of the reasoning engine proposed. The model captures a basic spatial structure of space using containment, adjacency and orientation relations. A spatial reasoning engine can utilise this structure to automatically derive implicit qualitative

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 689–694, 2008.

spatial relations between objects in a place ontology. The paper addresses one of the outstanding challenges in the domain of qualitative spatial reasoning, namely, combining quantitative and qualitative methods for geographic data management.

Some of the issues of concern in building place ontologies are presented in section 2. The proposed model of space is described in section 3 and pointers to how it supports reasoning are given in section 4.

## 2 Building Place Ontologies

Ontologies as knowledge repositories have been developed to support the primary goal of sharing knowledge in a manner that aids understanding. Geospatial ontologies [4] capture key conceptualisations of geographic domains to facilitate the reuse and sharing of geographic information on the web. Several examples of geo-ontology developments have recently been proposed [7,6].

Place ontologies are particular types of geospatial ontologies needed to support information retrieval activities on the web. These are models of terminology and structure of geographic space as well as records of entities in this space.

The web is viewed as a potentially rich resource for collecting place information to populate Place ontologies. Computational techniques are being developed to support the process of recognizing occurrences of place names in texts, and determining the corresponding geographical coordinates (e.g., [9,8,1]. Also, building place ontologies requires the integration of resources of variable quality in terms of precision and completeness.

Much of the semantics in Place ontologies are implicit and evident only at the instance level.For example, different types of spatial relationships exist between every object and all other objects in space; an object may be inside, north-of, near to, larger than another object, etc. Some of those relationships may be captured on the concept level but most others are implicit, evident only by visual interpretation and geometric computation. Explicit representation of such relationships is not practically possible and means for their automatic extraction are needed.

A principal research question is to identify which types and granularities of spatial relationships to model explicitly and which ones to compute or deduce. Much research has been undertaken in the area of representing spatial relationships and qualitative spatial reasoning [3,2,5] to support the composition of relationships. These methods can be adapted for representing and reasoning over the spatial structure of geospatial ontologies.

## 3 A Canonical Model of Spatial Relations

Three types of spatial relationships are most evident for geographic information retrieval on the web. These are containment (inside), proximity (close or near) and orientation (north-of, east-of, etc.) Here, we propose a minimal model that combines these three types of spatial relationships to support their effective manipulation. In particular, it is proposed that direct containment (parent-child relations), as well as direct adjacency (closest neighbours) and orientation relationships can provide a basic model to capture the implicit spatial structure of geospatial information.

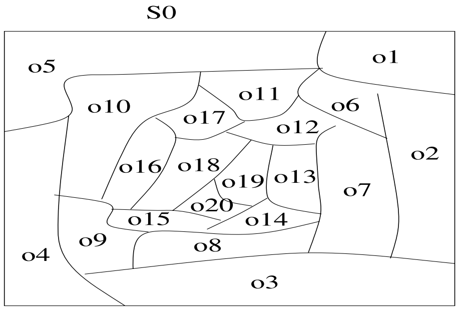
The basic idea of the proposed approach is simple. The map or spatial scene is divided up into concentric layers of adjacent objects in a form resembling the annual growth rings on a tree trunk. The map edges constitute the first ring and serves as a frame of reference for the rest of the scene. Objects are then grouped in a hierarchical division of the map into successive rings and related by explicitly defining adjacency and orientation relationships between objects in every ring. Adjacency and orientation relationships are then used to relate objects across consecutive rings. It is shown how the inter and intra-ring relationships can be represented using one structure.

Relationships are explicitly represented between a subset of objects on the map. It is shown how this information can be propagated to deduce orientation relations between non-adjacent objects. The representation levels are given in the rest of this section.

**I. The Ring Structure**

The semi-infinite region around the map edges is the reference ring and is bounded by the four map edges, north ($N$), east ($E$), south, ($S$) and west ($W$). Adjacent objects from the edge of the map form the first ring $R_1$. Adjacent objects to the first ring form the second ring $R_2$ and so on. This process ends when all the objects are related to a ring. The last ring formed is the core of the spatial scene. An example of this process is shown in figure 1 where successive rings are shown with no orientation or adjacency relationships.

Note that the virtual rings divides the containing objects into regions, namely $A_0, A_1, A_2$ and $A_3$, which are used to encode the adjacency and orientation



(a)

| $R_0$ | $N$ | $E$ | $S$ | $W$ | | |
|---|---|---|---|---|---|---|
| $R_1$ | $o_1$ | $o_2$ | $o_3$ | $o_4$ | $o_5$ | |
| $R_2$ | $o_6$ | $o_7$ | $o_8$ | $o_9$ | $o_{10}$ | $o_{11}$ |
| $R_3$ | $o_{12}$ | $o_{13}$ | $o_{14}$ | $o_{15}$ | $o_{16}$ | $o_{17}$ |
| $R_4$ | $o_{18}$ | $o_{19}$ | $o_{20}$ | | | |

(b)

**Fig. 1.** (a) Sample map scene. (b) Corresponding Rings.

relationships. Only objects with outer edges adjacent to the ring edges are related to it. In this case, the virtual rings act as frames of reference for their enclosing objects in exactly the same way the map edges are for the map scene.

**II. Representing Inter-Ring Relationships**
In this step, orientation relationships are represented between adjacent objects in different rings. For every object in every ring in the tree, define its relative orientation with adjacent objects. An example of encoding this type of relationships for the map in figure 2(a) is given in 2(b).

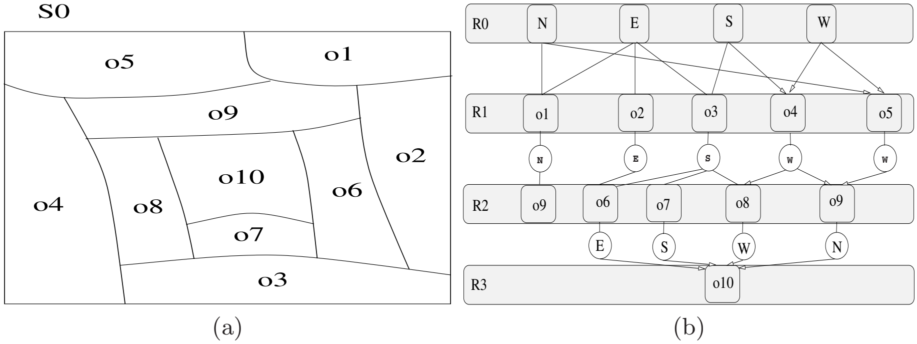


**Fig. 2.** Representing Inter-Ring relationships. (a) Sample map. (b) Orientation relationships are encoded only between adjacent objects.

**III. Representing Intra-Ring Relationships**
Here, orientation relationships are represented between adjacent objects in the same ring. This is achieved in two steps:

1. Firstly, a matrix is used to encode the adjacency relationships between objects in the ring. In figure 3(a) a matrix is used to hold adjacency relationships between objects in $R_2$. The fact that two entities are adjacent is represented by a value (1) in the matrix and by a value (0) otherwise. For example, $o_6$ is adjacent to both objects $o_7$ and $o_9$, object $o_7$ is adjacent to object $o_8$ and object $o_8$ is adjacent to object $o_9$. Since adjacency is a symmetric relation, the resulting matrix will be symmetric around the diagonal. Hence, only half the matrix is sufficient for the representation of the space topology and the matrix can be collapsed to the structure in figure 3(b).
2. Secondly, the values of (1) in the matrix is replaced by the relative orientation relationship between the corresponding objects as shown in figure 3(c). This structure is then denoted *Adjacency-Orientation Matrix*.

Note that $S_0$ represents the infinite embedding space of the map.

|  | $o_6$ | $o_7$ | $o_8$ | $o_9$ |
|---|---|---|---|---|
| $o_6$ | - | 1 | 0 | 1 |
| $o_7$ | 1 | - | 1 | 0 |
| $o_8$ | 0 | 1 | - | 1 |
| $o_9$ | 1 | 0 | 1 | - |

(a)

|  |  |  |  |
|---|---|---|---|
| $o_6$ |  |  |  |
| 1 | $o_7$ |  |  |
| 0 | 1 | $o_8$ |  |
| 1 | 0 | 1 | $o_9$ |

(b)

|  |  |  |  |
|---|---|---|---|
| $o_6$ |  |  |  |
| E | $o_7$ |  |  |
| 0 | E | $o_8$ |  |
| S | 0 | S | $o_9$ |

(c)

|  |  |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|
| $S_0$ |  |  |  |  |  |  |  |  |  |  |
| $E \wedge N$ | $o_1$ |  |  |  |  |  |  |  |  |  |
| $E$ | $N$ | $o_2$ |  |  |  |  |  |  |  |  |
| $E \wedge S$ | 0 | $N$ | $o_3$ |  |  |  |  |  |  |  |
| $S \wedge W$ | 0 | 0 | $E$ | $o_4$ |  |  |  |  |  |  |
| $W \wedge N$ | $E$ | 0 | 0 | $S$ | $o_5$ |  |  |  |  |  |
| 0 | 0 | $E$ | $S$ | 0 | 0 | $o_6$ |  |  |  |  |
| 0 | 0 | 0 | $S$ | 0 | 0 | $E$ | $o_7$ |  |  |  |
| 0 | 0 | 0 | $S$ | $W$ | 0 | 0 | $E$ | $o_8$ |  |  |
| 0 | $N$ | 0 | 0 | $W$ | 0 | $S$ | 0 | $S$ | $o_9$ |  |
| 0 | 0 | 0 | 0 | 0 | 0 | $E$ | $S$ | $W$ | $N$ | $o_{10}$ |

(d)

**Fig. 3.** (a) Adjacency matrix for objects in $R_2$ of figure 2. (b) Half the symmetric adjacency matrix is sufficient to capture the scene representation. (c) *Adjacency-orientation matrix* is formed by modifying the adjacency matrix to hold orientation relations between adjacent objects in the ring. (d) The combined inter- and intra- ring adjacency-orientation matrix.

**IV. Representing the Combined Inter- and Intra-Ring Relationships** Representation levels II and III above can be combined by representing the complete set of objects in the scene in the *adjacency-orientation matrix*. Hence, a uniform structure can be used to capture the adjacency and orientation relationships between all the map objects. The combined matrix for the map in figure 2 is shown in figure 2(d). The matrix can be kept compact by exploiting the transitive property of the orientation relations. Relations between non-adjacent entities can be deduced using qualitative reasoning. The convention of orientation relations is $R(column, row)$.

## 4 Reasoning with the Ring Model

The structure proposed supports the automatic derivation of implicit relationships as follows.

1. Orientation relationships between consecutive objects on every ring edge are determined by their order. For example, east and west edges in every ring determine the north and south relationships, e.g. in figure 2(a), object 1 is north of object 2 which is north of 3, etc. Similarly, the north and south edges determine west and east relationships.
2. Transitivity of order relations are used to determine orientation relations between objects on similar orientation edges in consecutive rings. If $A$ is on the east edge of ring 0 and B and C are on the east edges of rings 1 and 2 respectively, then, $east(A, B) \wedge east(B, C) \rightarrow east(A, C)$.
3. Projections of ring edges as well as the order of the ring can be used to determine further refined orientation relationships across rings.
4. The implicit order of the rings can be used to imply qualitative distance relationships, where objects in consecutive rings are closer to each others that those in subsequent rings.

In addition to the utilisation of the model to derive a basic structure of space, qualitative spatial reasoning techniques can be applied to propagate relationships across networks of objects and to define spatial integrity constraints to maintain the consistency of the stored and derived structures.

## 5 Conclusions

In this paper, a qualitative model is proposed to capture the basic structure of spatial scenes. The model encodes some adjacency and orientation relationships between objects and supports the automatic derivation of others. An algorithm for systematically defining rings of orientation relations is given with which relations between adjacent and non-adjacent objects could be accurately defined using a minimal set of pre-stored and calculated information. The method has the advantage of minimising the uncertainty associated with chain reasoning when applied in this domain.

The model is proposed as a base to support the building and management of large place ontologies used in the context of geographic information retrieval on the web. It provides a qualitative structure of space that complements traditional quantitative computational geometry techniques and supports the practical development of spatial reasoning engines to manage large geospatial data stores. Such a combined approach is particularly useful on the web where resources for precise place information are limited.

## References

1. Arampatzis, A., Van Kreveld, M., Reinbacher, I., Jones, C., Vaid, S., Joho, H., Clough, P., Sanderson, M.: Web-based delineation of imprecise regions. Computers, Environment and Urban Systems 30(4), 436–459 (2006)
2. Chang, S., Shi, Q., Yan, W.: Iconic Indexing by 2-D Strings. IEEE Trans. Pattern Analysis and Machine Intelligence PAMI9(3), 413–428 (1987)
3. Cohn, A.G., Renz, J.: Qualitative spatial reasoning. In: van Harmelen, F., Lifschitz, V., Porter, B. (eds.) Handbook of Knowledge Representation. Elsevier, Amsterdam (2007)
4. Egenhofer, M.J.: Toward the semantic geospatial web. In: Proceedings of the tenth ACM international symposium on Advances in geographic information systems, pp. 1–4. ACM Press, New York (2002)
5. Glasgow, J., Papadias, D.: Computation Imagery. In: Chandrasekaran, B., Glasgow, J. (eds.) Diagrammatic Reasoning, pp. 435–480. AAAI Press, Menlo Park (1995)
6. Hiramatsu, K., Reitsma, F.: Georeferencing the semantic web: ontology based markup of geographically referenced information. In: Joint EuroSDR/EuroGeographics workshop (2004)
7. Jones, C., Abdelmoty, A., Fu, G.: Maintaining ontologies for geographical information retrieval on the web. In: Meersman, R., Tari, Z., Schmidt, D.C. (eds.) CoopIS 2003, DOA 2003, and ODBASE 2003. LNCS, vol. 2888, pp. 934–951. Springer, Heidelberg (2003)
8. Leidnes, K., Sinclair, G., Webber, B.: Grounding spatial named entities for information extraction and question answering. In: TLNAACL Workshop on Analysis of Geographic References 2003, pp. 31–38 (2003)
9. Silva, M., Martins, B., Chaves, M., Afonso, A., Cardoso, N.: Adding geographic scopes to web resources. Computers, Environment and Urban Systems 30(4), 378–399 (2006)

# Towards Relational Inconsistent Databases with Functional Dependencies

Sergio Greco and Cristian Molinaro

DEIS, Univ. della Calabria, 87036 Rende, Italy
{greco,cmolinaro}@deis.unical.it

**Abstract.** This paper investigates the problem of repairing inconsistent databases in the presence of functional dependencies. Specifically, we present a repairing strategy where only tuple updates are allowed in order to restore consistency. The proposed approach allows us to obtain a unique repair which can be computed in polynomial time.

## 1 Introduction

Inconsistent databases, namely databases which violate given integrity constraints, may arise in several scenarios, such as database integration, data warehousing, automated reasoning systems and others. In the presence of an inconsistent database two possible solutions have been investigated in the literature: (i) repairing the database or (ii) computing consistent answers over the inconsistent database [2,3,7,8,12,14,15,16,20]. Let's intuitively clarify the notions of repair and consistent answer. A *repair* (for a database and a set of integrity constraints) is a set of update operations which lead the database to a consistent state (*repaired database*) by preserving its data as much as possible. In most of the approaches proposed in literature, only insertions and deletions of tuples are allowed as update operations. The notion of *consistent answer* over an inconsistent database has been firstly given in [7]. Intuitively, consistent answers are those tuples which can be derived from all the repaired databases. The problem with such a semantics is that computing consistent answers is co-$\mathcal{NP}$-hard also for simple classes of integrity constraints such as functional dependencies or even primary keys [11,17].

In this paper we propose a repairing strategy which consists in updating the database by means of *attribute modifications* in the case of functional dependency violations . No deletion operation is used to restore consistency.

In order to show this approach, let's consider the following example.

*Example 1.* Consider the following database DB

Employee

| Name | Level | Salary |
|---|---|---|
| Mary | A | 20 |
| John | A | 40 |
| Peter | B | 30 |

and the functional dependency fd : Level → Salary defined over Employee. The database DB is inconsistent as it violates fd: there are two different employees (Mary and

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 695–702, 2008.

John) having the same Level, but different Salary. In this case the repaired database $\mathtt{DB^R}$ consists of the following relation $\mathtt{Employee^R}$[1]:

$\mathtt{Employee^R}$

| Name | Level | Salary |
|---|---|---|
| Mary | A | #1 |
| John | A | #1 |
| Peter | B | 30 |

Therefore, in order to satisfy fd, the inconsistent tuples in the relation Employee are updated. Specifically, all the tuples having the value of the attribute Level equal to A, take for the attribute Salary the same unknown value #1 belonging to the set $\{20, 40\}$. □

As it will be formally shown in the paper, given an inconsistent database and a set of functional dependencies, the proposed repairing strategy allow us to obtain a unique repair (modulo renaming of the new unknown values).

## 2 Preliminaries

This section introduces preliminaries on relational databases and integrity constraints ([1,24]).

### 2.1 Relational Databases

The existence of *alphabets* of *relation symbols* (or *predicate symbols*) and *attribute symbols* is assumed. The *domain* of an attribute A is denoted by Dom(A). The *database domain* is denoted by Dom.

A *relation schema* S is of the form $\mathtt{p(A_1, ..., A_m)}$ where p is a relation symbol and $\mathtt{A_i}$ ($\mathtt{i} \in [\mathtt{1..m}]$) is an attribute symbol. A *relation instance* (or simply *relation*) r over S is any subset of $\mathtt{Dom(A_1)} \times \cdots \times \mathtt{Dom(A_m)}$. A *tuple* is each element of r.

A *database schema* DS is of the form $\langle \mathtt{RS}, \mathtt{IC} \rangle$ where $\mathtt{RS} = \{\mathtt{S_1}, \ldots, \mathtt{S_n}\}$ is a set of relation schemata and IC is a set of *integrity constraints*. A *database instance* (or simply *database*) DB over DS is a set $\{\mathtt{r_1}, \ldots, \mathtt{r_n}\}$ of relations over $\{\mathtt{S_1}, \ldots, \mathtt{S_n}\}$ (See [19] for a more comprehensive introduction of the relational model of data).

Given a database schema $\mathtt{DS} = \langle \mathtt{RS}, \mathtt{IC} \rangle$ and a database instance DB over DS, we say that DB is *consistent* if $\mathtt{DB} \models \mathtt{IC}$, i.e. if all the integrity constraints in IC are satisfied by DB, otherwise it is *inconsistent*.

In the following we can refer to a tuple $\langle \mathtt{u_1}, \ldots, \mathtt{u_n} \rangle$ of a relation r with the ground atom $\mathtt{r(u_1, \ldots, u_n)}$. An *atom* is of the form $\mathtt{q(t_1, \ldots, t_n)}$ where q is a relation symbol and $\mathtt{t_1, \ldots, t_n}$ are terms, i.e. variables or constants.

[1] Note that, by adopting the classical repairing strategy (consisting in tuple insertions/deletions) there are two different repaired databases obtained by deleting one of the two employee tuples whose value on the attribute Level is A.

### 2.2 Integrity Constraints

Databases contain, other than data, intentional knowledge expressed by means of integrity constraints. Integrity constraints express semantic information over data, i.e. relationships that should hold among data and they are mainly used to validate database transactions. They are usually defined by first order rules or by means of special notations used for particular classes of them (such as functional dependencies).

**Definition 1.** An *integrity constraint* is a formula of the first order predicate calculus of the form[2]:

$$(\forall\, \mathtt{X})[\bigwedge_{\mathtt{i=1}}^{\mathtt{m}} \mathtt{b_i}(\mathtt{X_i}), \Phi(\mathtt{X_0}) \supset \bigvee_{\mathtt{i=m+1}}^{\mathtt{n}} (\exists\, \mathtt{Z_i})\, \mathtt{b_i}(\mathtt{X_i}, \mathtt{Z_i})]$$

where $\mathtt{b_i}$ ($\mathtt{i} \in [1..\mathtt{n}]$) is a predicate symbol, $\Phi(\mathtt{X_0})$ denotes a conjunction of built-in atoms, $\mathtt{X} = \bigcup_{\mathtt{i=1}}^{\mathtt{m}} \mathtt{X_i}$, $\mathtt{X_j} \subseteq \mathtt{X}$ ($\mathtt{j} \in [0..\mathtt{n}]$) and all existentially quantified variables appear once. □

In the definition above, the conjunction $\bigwedge_{\mathtt{i=1}}^{\mathtt{m}} \mathtt{b_i}(\mathtt{X_i}), \Phi(\mathtt{X_0})$ is called *body* and the disjunction $\bigvee_{\mathtt{i=m+1}}^{\mathtt{n}} (\exists\, \mathtt{Z_i})\, \mathtt{b_i}(\mathtt{X_i}, \mathtt{Z_i})$ *head* of the integrity constraints.

*Example 2.* The integrity constraint $(\forall\, \mathtt{X})\ [\, \mathtt{p(X)} \supset \mathtt{q(X)} \vee \mathtt{r(X)}\, ]$ states that the relation p must be contained in the union of the relations q and r. □

In the following we concentrate our attention to a particular class of integrity constraints consisting of functional dependencies.

**Definition 2.** A *functional dependency* fd is a constraint of the form

$$\forall(\mathtt{X,Y,U,Z,V})\ [\, \mathtt{p(X,Y,U)},\ \mathtt{p(X,Z,V)} \supset \mathtt{Y} = \mathtt{Z}] \quad (1)$$

where $\mathtt{X,Y,U,Z,V}$ are lists of variables. A database DB satisfies fd if for each $\mathtt{p(x,y,u)}$, $\mathtt{p(x,z,v)} \in \mathtt{DB}$, then $\mathtt{y} = \mathtt{z}$. □

## 3 Repairing Inconsistent Databases

Most of the classical approaches to the problem of repairing inconsistent databases rely on minimal sets of insert/delete operations; only few works have investigated the computation of repairs consisting of tuple updates ([25]).

In the following, we present a technique for repairing databases in the presence of functional dependencies; such a technique relies on tuple updates.

Before formally introducing the notion of repair, let us introduce some preliminary definitions.

We assume to have an infinite enumerable domain $\mathtt{D\#} = \{\mathtt{\#1}, ..., \mathtt{\#n}, ...\}$ of distinct *unknown values*. We denote with $\mathtt{Dom}^{\#} = \mathtt{Dom} \cup \mathtt{D\#}$. Every relation over the

[2] The order of literals in a conjunction or in a disjunction is immaterial. A literal can appear in a conjunction or in a disjunction at most once. The meaning of the symbols ‘∧’ and ‘,’ is the same.

schema $\mathtt{r}(\mathtt{A}_1,...,\mathtt{A}_n)$ can take values from $\mathtt{Dom}^{\#}(\mathtt{A}_1)\times\cdots\times\mathtt{Dom}^{\#}(\mathtt{A}_n)$, where $\mathtt{Dom}^{\#}(\mathtt{A}_i) = \mathtt{Dom}(\mathtt{A}_i)\cup\mathtt{D\#}$.

Unknown values are introduced to replace conflicting values. Thus, given a relation $\mathtt{r}$ with schema $\mathtt{r}(\mathtt{A}_1,...,\mathtt{A}_n)$ and a tuple $\mathtt{t} = \mathtt{r}(\mathtt{u}_1,...,\mathtt{u}_n)$, a value $\mathtt{u}_i = \mathtt{\#i}$ means that the tuple $\mathtt{t}$ has been introduced into the relation $\mathtt{r}$ to replace conflicting tuples and that the value of the $\mathtt{i}$-th attribute could be any value associated with the $\mathtt{i}$-th attribute of the conflicting tuples. In the following, for each value $\mathtt{u}_i$ of a tuple $\mathtt{t}$, $\mathtt{dom}(\mathtt{u}_i)$ denotes the set of values which can be associated with the $\mathtt{i}$-th attribute of $\mathtt{t}$.

Let's now clarify how the proposed approach works in order to restore consistency if a violation of a functional dependency occurs.

Consider a functional dependency $\mathtt{fd}$ of the form (1). If the database $\mathtt{DB}$ contains a pair of ground atoms $\mathtt{p}(\mathtt{a},\mathtt{b}_1,\mathtt{c}_1)$ and $\mathtt{p}(\mathtt{a},\mathtt{b}_2,\mathtt{c}_2)$, then the functional dependency $\mathtt{fd}$ is violated; moreover it could be satisfied by updating both the atoms to $\mathtt{p}(\mathtt{a},\mathtt{\#i},\mathtt{c}_1)$ and $\mathtt{p}(\mathtt{a},\mathtt{\#i},\mathtt{c}_2)$, where $\mathtt{\#i}$ is a new unknown value and $\mathtt{dom}(\mathtt{\#i}) = \{\mathtt{b}_1,\mathtt{b}_2\}$ (i.e. the value of $\mathtt{\#i}$ could be either $\mathtt{b}_1$ or $\mathtt{b}_2$).

Before formally introducing the repairing technique, let us introduce the satisfaction of functional dependencies for databases whose tuples can take values from $\mathtt{Dom}^{\#}$.

**Definition 3.** Given a relation $\mathtt{r}$ over the schema $\mathtt{r}(\mathtt{A}_1,...,\mathtt{A}_n)$ and a functional dependency $\mathtt{fd} = \mathtt{X}\rightarrow\mathtt{A}$, where $\mathtt{X}$ and $\mathtt{A}$ are a set of attributes and an attribute respectively, then $\mathtt{r}\models\mathtt{fd}$ if $\forall\mathtt{t}_1,\mathtt{t}_2\in\mathtt{r}$, $\bigwedge_{\mathtt{A}_i\in\mathtt{X}}\mathtt{dom}(\mathtt{t}_1(\mathtt{A}_i))\cap\mathtt{dom}(\mathtt{t}_2(\mathtt{A}_i))\neq\emptyset$ implies $\mathtt{t}_1(\mathtt{A}) = \mathtt{t}_2(\mathtt{A})$. □

It is worth noting that the previous definition extends the classical definition of satisfiability of functional dependencies over standard databases. Let us now discuss how databases can be repaired to satisfy functional dependencies.

**Definition 4.** Given a relation $\mathtt{r}$ over $\mathtt{r}(\mathtt{A}_1,...,\mathtt{A}_n)$ and two tuples $\mathtt{t}_1 = \mathtt{r}(\mathtt{u}_1,...,\mathtt{u}_n)$ and $\mathtt{t}_2 = \mathtt{r}(\mathtt{v}_1,...,\mathtt{v}_n)$ we say that $\mathtt{t}_1$ *subsumes* $\mathtt{t}_2$ (denoted by $\mathtt{t}_1\gtrapprox\mathtt{t}_2$) if for all $\mathtt{i}\in[1..\mathtt{n}]$ $\mathtt{dom}(\mathtt{v}_i)\subseteq\mathtt{dom}(\mathtt{u}_i)$. Moreover, $\mathtt{t}_1$ *strictly subsumes* $\mathtt{t}_2$ (denoted by $\mathtt{t}_1\gtrsim\mathtt{t}_2$) if $\mathtt{t}_1\gtrapprox\mathtt{t}_2$ and $\mathtt{t}_2\not\gtrapprox\mathtt{t}_1$. □

*Example 3.* Consider the tuples $\mathtt{t}_1 = \mathtt{r}(\mathtt{a},\mathtt{\#1},\mathtt{c})$ and $\mathtt{t}_2 = \mathtt{r}(\mathtt{a},\mathtt{b},\mathtt{c})$ with $\mathtt{dom}(\mathtt{\#1}) = \{\mathtt{a},\mathtt{b},\mathtt{c}\}$; then $\mathtt{t}_1\gtrsim\mathtt{t}_2$. □

**Definition 5.** A (*tuple*) *substitution* $\mathtt{S}$ is a set of pairs of tuples $\{\mathtt{t}_1/\mathtt{t}'_1,...,\mathtt{t}_n/\mathtt{t}'_n\}$, where for all $\mathtt{i},\mathtt{j}\in[1..\mathtt{n}]$, with $\mathtt{i}\neq\mathtt{j}$, the conditions $\mathtt{t}_i\neq\mathtt{t}_j$ and $\mathtt{t}_i\lessapprox\mathtt{t}'_i$ hold. □

Given a substitution $\mathtt{S} = \{\mathtt{A}_1/\mathtt{B}_1,...,\mathtt{A}_n/\mathtt{B}_n\}$, we denote with $\mathtt{S}[1]$ and $\mathtt{S}[2]$ the left and right sides of $\mathtt{S}$, respectively, i.e. $\mathtt{S}[1] = \{\mathtt{A}_1,...,\mathtt{A}_n\}$ whereas $\mathtt{S}[2] = \{\mathtt{B}_1,...,\mathtt{B}_n\}$.

**Definition 6.** Let $\mathtt{S} = \{\mathtt{A}_1/\mathtt{B}_1,...,\mathtt{A}_n/\mathtt{B}_n\}$ and $\mathtt{T} = \{\mathtt{C}_1/\mathtt{D}_1,...,\mathtt{C}_m/\mathtt{D}_m\}$ be two substitutions, then $\mathtt{T}$ *subsumes* $\mathtt{S}$ (denoted by $\mathtt{T}\gtrapprox\mathtt{S}$) if $\mathtt{S}[1]\subseteq\mathtt{T}[1]$ and for each pair $\mathtt{A}_i,\mathtt{C}_j,\ \mathtt{A}_i = \mathtt{C}_j$ implies $\mathtt{D}_j\gtrapprox\mathtt{B}_i$. Moreover, $\mathtt{T}$ *strictly subsumes* $\mathtt{S}$ (denoted by $\mathtt{T}\gtrsim\mathtt{S}$) if $T\gtrapprox S$ and $S\not\gtrapprox T$. □

Therefore, given a database $\mathtt{DB}$ and a substitution $S$, the application of $S$ to $\mathtt{DB}$, denoted by $\mathtt{DB}\circ\mathtt{S}$, gives the following database $\{\mathtt{t}|\mathtt{t}\in\mathtt{DB}\wedge\mathtt{t}\notin\mathtt{S}[1]\}\cup\{\mathtt{t}'|\exists\mathtt{t}\in\mathtt{DB}\wedge\mathtt{t}/\mathtt{t}'\in\mathtt{S}\}$, that is $\mathtt{DB} \circ \mathtt{S}$ contains the tuples in DB which are not substituted plus the new tuples which are entered to replace old tuples.

Let us now introduce the formal definition of database repair w.r.t. a set of functional dependencies.

**Definition 7.** *(FD_Repair).* Given a database DB and a set FD of functional dependencies, a repair for $\langle \mathtt{DB}, \mathtt{FD} \rangle$ is a substitution R such that:

1. $\mathtt{DB} \circ \mathtt{R} \models \mathtt{FD}$,
2. there not exists a substitution S such that
   - $\mathtt{DB} \circ \mathtt{S} \models \mathtt{FD}$, and
   - $\mathtt{S} \lesssim \mathtt{R}$.
3. each tuple $\mathtt{t} \in \mathtt{DB}$ can be obtained from a tuple in $\mathtt{DB} \circ \mathtt{R}$ by replacing unknown values with the associated constants. □

In the definition above Item 1 states that the database obtained by applying R on the source database is consistent whereas Item 2 verifies that R is "minimal".

*Example 4.* Consider the database schema consisting of the relation $\mathtt{emp}(\mathtt{name}, \mathtt{city}, \mathtt{dept}, \mathtt{mgr})$ with the following set IC of functional dependencies:

- $\mathtt{fd}_1 = \mathtt{name} \rightarrow \mathtt{city}$ stating that an employee has to work in a unique city;
- $\mathtt{fd}_2 = \mathtt{dept} \rightarrow \mathtt{city}$ and $\mathtt{fd}_3 = \mathtt{dept} \rightarrow \mathtt{mgr}$ stating that a department must be located in a unique city and have a unique manager.

Let DB be the following instance:

$$\begin{array}{l}\mathtt{emp}(\mathtt{john}, \mathtt{NY}, \mathtt{admin}, \mathtt{frank}),\\ \mathtt{emp}(\mathtt{john}, \mathtt{WA}, \mathtt{sales}, \mathtt{daniel}),\\ \mathtt{emp}(\mathtt{mary}, \mathtt{LA}, \mathtt{sales}, \mathtt{susan}),\\ \mathtt{emp}(\mathtt{julie}, \mathtt{NY}, \mathtt{admin}, \mathtt{frank})\,.\end{array}$$

A repair R for $\langle \mathtt{DB}, \mathtt{IC} \rangle$ consists of the following substitution:

$$\begin{array}{l}\mathtt{emp}(\mathtt{john}, \mathtt{NY}, \mathtt{admin}, \mathtt{frank})/\mathtt{emp}(\mathtt{john}, \#1, \mathtt{admin}, \mathtt{frank}),\\ \mathtt{emp}(\mathtt{john}, \mathtt{WA}, \mathtt{sales}, \mathtt{daniel})/\mathtt{emp}(\mathtt{john}, \#1, \mathtt{sales}, \#2),\\ \mathtt{emp}(\mathtt{mary}, \mathtt{LA}, \mathtt{sales}, \mathtt{susan})/\mathtt{emp}(\mathtt{mary}, \#1, \mathtt{sales}, \#2).\\ \mathtt{emp}(\mathtt{julie}, \mathtt{NY}, \mathtt{admin}, \mathtt{frank})/\mathtt{emp}(\mathtt{julie}, \#1, \mathtt{admin}, \mathtt{frank}).\end{array}$$

The corresponding repaired database $\mathtt{DB}^{\mathtt{R}}$ consists of the following tuples:

$$\begin{array}{l}\mathtt{emp}(\mathtt{john}, \#1, \mathtt{admin}, \mathtt{frank}),\\ \mathtt{emp}(\mathtt{john}, \#1, \mathtt{sales}, \#2),\\ \mathtt{emp}(\mathtt{mary}, \#1, \mathtt{sales}, \#2),\\ \mathtt{emp}(\mathtt{julie}, \#1, \mathtt{admin}, \mathtt{frank})\end{array}$$

where $\mathtt{dom}(\#1) = \{\mathtt{NY}, \mathtt{WA}, \mathtt{LA}\}$ and $\mathtt{dom}(\#2) = \{\mathtt{daniel}, \mathtt{susan}\}$. □

Observe that in the example above any substitution obtained from R by replacing #1 and #2 with two distinct constants, respectively in $\mathtt{dom}(\#1)$ and $\mathtt{dom}(\#2)$, is a repair as well.

**Theorem 1.** *Given a database* $\mathtt{DB}$ *and a set* $\mathtt{FD}$ *of functional dependencies, there exists a unique substitution, modulo renaming of values in* $\mathtt{D\#}$*, which is a repair for* $\langle \mathtt{DB}, \mathtt{FD} \rangle$.

**Proof.** Assume that there are two repairs $\mathtt{R}$ and $\mathtt{S}$. Clearly, both $\mathtt{R}$ and $\mathtt{S}$ are applicable and satisfy $\mathtt{IC}$. In order to have two repairs we must have that i) $\mathtt{R}[1] \not\subset \mathtt{S}[1] \ \wedge \ \mathtt{S}[1] \not\subset \mathtt{R}[1]$, and ii) $\mathtt{R} \not\lesssim \mathtt{S} \ \wedge \ \mathtt{S} \not\lesssim \mathtt{R}$.

Concerning Condition i) we have that $\mathtt{S}[1] = \mathtt{R}[1]$ as the set of conflicting tuple is not determined by the repair and every repair must replace all conflicting tuples.

Concerning Condition ii) in order to have two different repair $\mathtt{R}$ and $\mathtt{S}$ there must be a tuple $\mathtt{t} = \mathtt{p}(\mathtt{x}, \mathtt{y}, \mathtt{z})$ conflicting with some other tuple $\mathtt{t}' = \mathtt{p}(\mathtt{x}, \mathtt{y}', \mathtt{z}')$ on the base of a functional dependency $\mathtt{fd} = \mathtt{X} \rightarrow \mathtt{Z}$, from the attributes $\mathtt{X}$ corresponding to the value $\mathtt{x}$ to the attribute $\mathtt{Z}$ corresponding to the value $\mathtt{z}$. Anyhow, the tuple $\mathtt{t}$ cannot be replaced by a tuple $\mathtt{t_i} = \mathtt{p}(\mathtt{x}, \mathtt{y}, \mathtt{\#i})$ in $\mathtt{R}$ and by a tuple $\mathtt{t_j} = \mathtt{p}(\mathtt{x}, \mathtt{y}, \mathtt{\#j})$ is $\mathtt{S}$ with $\mathtt{\#i} \neq \mathtt{\#j}$ as both $\mathtt{\#i}$ and $\mathtt{\#j}$ are associated with the same domain. □

The next algorithm shows how to compute the unique repair (modulo renaming of values in $\mathtt{D\#}$) for a database and a set of functional dependencies.

**Algorithm 1. Computing FD_Repair**
**Input:** a database $\mathtt{DB}$ and a set $\mathtt{FD}$ of functional dependencies.
**Output:** a repair $\mathtt{R}$ for $\langle \mathtt{DB}, \mathtt{IC} \rangle$

```
begin
    $R = \emptyset;\ DB_0 = DB;\ i = 0;$
    while ($\exists fd = A_1 ... A_n \rightarrow A_0 \in FD$ s.t. $DB \not\models fd$) do begin
        $S = \emptyset;\ \ i := i + 1;$
        Let $T$ be a maximal set of tuples conflicting each other w.r.t. $fd$;
        Let $U = \{\#j \mid \exists t \in T \wedge t(A_0) = \#j\}$
        Generate new value $\#i$ such that $dom(\#i) = \cup_{t \in T} dom(t(A_0))$;
        $S = \{t/t' \mid t \in T \wedge t'$ is obtained from $t$ by replacing $t(A_0)$ with $\#i\} \cup$
            $\{t/t' \mid t \in DB \wedge t$ contains some $\#j \in U \wedge$
             $t'$ is obtained from $t$ by replacing each $\#j \in U$ with $\#i\}$;
        $DB := DB \circ S;$
        $R = \{t/t' \in R \circ S$ s.t. $t \in DB_0\}$;
    end
end
```

Observe that the algorithm above uses the standard concepts of application and composition of substitutions ([**?**]).

*Example 5.* Consider the following database $\mathtt{DB} = \{\mathtt{r}(\mathtt{a_1}, \mathtt{b_1}), \mathtt{r}(\mathtt{a_1}, \mathtt{b_2}), \mathtt{r}(\mathtt{a_2}, \mathtt{b_2})\}$ over the schema $\mathtt{r}(\mathtt{A}, \mathtt{B})$ and the functional dependencies $\mathtt{f_1} = \mathtt{A} \rightarrow \mathtt{B}$ and $\mathtt{f_2} = \mathtt{B} \rightarrow \mathtt{A}$.

Initially $R = \emptyset$.

As $\mathtt{DB} \not\models \mathtt{FD}$ the while loop is executed. Assume that at the first iteration the functional dependency $\mathtt{f_1}$ is selected. Then, the substitution $S = \{\ \mathtt{r}(\mathtt{a_1}, \mathtt{b_1})/\mathtt{r}(\mathtt{a_1}, \mathtt{\#1}), \mathtt{r}(\mathtt{a_1}, \mathtt{b_2})/\mathtt{r}(\mathtt{a_1}, \mathtt{\#1})\ \}$ with $\mathtt{dom}(\mathtt{\#1}) = \{\mathtt{b_1}, \mathtt{b_2}\}$ is computed. By applying $S$ to $\mathtt{DB}$ and composing it with $R$, we get the database $\mathtt{DB} = \{\mathtt{r}(\mathtt{a_1}, \mathtt{\#1}), \mathtt{r}(\mathtt{a_2}, \mathtt{b_2})\}$ and the substitution $R = \{\ \mathtt{r}(\mathtt{a_1}, \mathtt{b_1})/\mathtt{r}(\mathtt{a_1}, \mathtt{\#1}),\ \mathtt{r}(\mathtt{a_1}, \mathtt{b_2})/\mathtt{r}(\mathtt{a_1}, \mathtt{\#1})\ \}$.

At the second iteration $\mathtt{DB} \not\models \mathtt{f_2}$. Then, the substitution $S = \{\mathtt{r(a_2, b_2)/r(\#2, b_2)}, \mathtt{r(a_1, \#1)/r(\#2, \#1)}\}$ with $\mathtt{dom(\#2)} = \{\mathtt{a_1, a_2}\}$ is computed. By applying $S$ to DB and composing it with $R$, we get the database $\mathtt{DB} = \{\mathtt{r(\#2, \#1), r(\#2, b_2)}\}$ and the substitution $R = \{\mathtt{r(a_1, b_1)/r(\#2, \#1), r(a_1, b_2)/r(\#2, \#1), r(a_2, b_2)/r(\#2, b_2)}\}$.

At the third iteration $\mathtt{DB} \not\models \mathtt{f_1}$. So, the substitution $S = \{\mathtt{r(\#2, \#1)/r(\#2, \#3)}, \mathtt{r(\#2, b_2)/r(\#2, \#3)}\}$ with $\mathtt{dom(\#3)} = \{\mathtt{b_1, b_2}\}$ is computed. By applying $S$ to DB and composing it with $R$, we get the database $\mathtt{DB} = \{\mathtt{r(\#2, \#3)}\}$ and the substitution $R = \{\mathtt{r(a_1, b_1)/r(\#2, \#3), r(a_1, b_2)/r(\#2, \#3), r(a_2, b_2)/r(\#2, \#3)}\}$.

At this point DB is consistent. □

**Theorem 2.** *Given a database* DB *and set* FD *of functional dependencies, Algorithm 1 computes the unique repair (modulo renaming of constants in* D#*) for* $\langle \mathtt{DB}, \mathtt{FD} \rangle$. □

**Proposition 1.** *Given a database* DB *and a set* FD *of functional dependencies, the complexity of computing a repaired database is polynomial time.* □

## References

1. Abiteboul, S., Hull, R., Vianu, V.: Foundations of Databases. Addison-Wesley, Reading (1994)
2. Agarwal, S., Keller, A.M., Wiederhold, G., Saraswat, K.: Flexible Relation: an Approach for Integrating Data from Multiple, Possibly Inconsistent Databases. In: ICDE (1995)
3. Arenas, M., Bertossi, L., Chomicki, J.: Consistent Query Answers in Inconsistent Databases. In: Proc. PODS 1999, pp. 68–79 (1999)
4. Arenas, M., Bertossi, L., Chomicki, J.: Specifying and Querying Database repairs using Logic Programs with Exceptions. In: FQAS Conf., pp. 27–41 (2000)
5. Bravo, L., Bertossi, L.: Semantically Correct Query Answers in the Presence of Null Values. In: Grust, T., Höpfner, H., Illarramendi, A., Jablonski, S., Mesiti, M., Müller, S., Patranjan, P.-L., Sattler, K.-U., Spiliopoulou, M., Wijsen, J. (eds.) EDBT 2006. LNCS, vol. 4254, pp. 336–357. Springer, Heidelberg (2006)
6. Bertossi, L., Chomicki, J.: Query Answering in Inconsistent Databases. Logics for Emerging Applications of Databases, 43–83 (2003)
7. Bry, F.: Query Answering in Information System with Integrity Constraints. In: IICIS, pp. 113–130 (1997)
8. Cali, A., Calvanese, D., De Giacomo, G., Lenzerini, M.: Data Integration under Integrity Constraints. In: Pidduck, A.B., Mylopoulos, J., Woo, C.C., Ozsu, M.T. (eds.) CAiSE 2002. LNCS, vol. 2348, pp. 262–279. Springer, Heidelberg (2002)
9. Cali, A., Lembo, D., Rosati, R.: On the Decidability and Complexity of Query Answering over Inconsistent and Incomplete Databases. In: PODS, pp. 260–271 (2003)
10. Caroprese, L., Zumpano, E.: A Framework for Merging, Repairing and Querying Inconsistent Databases. In: ADBIS, pp. 383–398 (2006)
11. Chomicki, J., Marcinkowski, J.: Minimal-change integrity maintenance using tuple deletions. Information & Computation 197(1/2), 90–121 (2005)
12. Dung, P.M.: Integrating Data from Possibly Inconsistent Databases. In: COOPIS, pp. 58–65 (1996)
13. Fagin, R., Kolaitis, P.G., Miller, R.J., Popa, L.: Data exchange: semantics and query answering. Theor. Comput. Sci. 336(1), 89–124 (2005)
14. Grant, J., Subrahmanian, V.S.: Reasoning in Inconsistent Knowledge Bases. IEEE-TKDE 7(1), 177–189 (1995)

15. Greco, S., Zumpano, E.: Querying Inconsistent Database. In: Parigot, M., Voronkov, A. (eds.) LPAR 2000. LNCS (LNAI), vol. 1955, pp. 308–325. Springer, Heidelberg (2000)
16. Greco, G., Greco, S., Zumpano, E.: A Logic Programming Approach to the Integration, Repairing and Querying of Inconsistent Databases. In: Codognet, P. (ed.) ICLP 2001. LNCS, vol. 2237, pp. 348–364. Springer, Heidelberg (2001)
17. Greco, G., Greco, S., Zumpano, E.: A Logical Framework for Querying and Repairing Inconsistent Databases. TKDE 15(6), 1389–1408 (2003)
18. Imielinski, T.: Incomplete Information in Logical Databases. IEEE Data Eng. Bull. 12(2), 29–40 (1989)
19. Kanellakis, P.C.: Elements of Relational Database Theory. In: van Leewen, J. (ed.) Handbook of Theoretical Computer Science, vol. 2. North-Holland, Amsterdam (1991)
20. Lin, J.: A Semantics for Reasoning Consistently in the Presence of Inconsistency. AI 86(1), 75–95 (1996)
21. Lloyd, J.: Foundation of Logic Programming. Springer, Heidelberg (1987)
22. Reiter, R.: A sound and sometimes complete query evaluation algorithm for relational databases with null values. J. ACM 33(2), 349–370 (1986)
23. Yan, L.L., Ozsu, M.T.: Conflict Tolerant Queries in Aurora Coopis, pp. 279–290 (1999)
24. Ullman, J.D.: Principles of Database and Knowledge-Base Systems, vol. 1. Computer Science Press (1998)
25. Wijsen, J.: Database repairing using updates. ACM Transactions Database Systems 30(3), 722–768 (2005)

# A Bank Run Model in Financial Crises

Katsutoshi Yada[1], Takashi Washio[2], Yasuharu Ukai[1], and Hisao Nagaoka[1]

[1] Research Center of Socionetwork Strategies, Kansai University, 3-3-35, Yamate, Suita,
564-8680 Osaka, Japan
{yada, ukai, nagaoka}@kansai-u.ac.jp
[2] ISIR, Osaka University, 8-1, Mihogaoka, Ibaraki,
567-0047 Osaka, Japan
washio@ar.sanken.osaka-u.ac.jp

**Abstract.** Financial crises are occurring frequently in Asia, Europe, and America, and it is important for banks to investigate strategies for such crises. The objective of this research is to build a model for bank runs by deposit holders in financial crises, and use that model to present a framework for estimating the amounts of deposit withdrawals during financial crises. By carrying out detailed investigation of the bank run model thus constructed, we clarified that the characteristics of customers and branch locations both bring about differences in bank runs. Our estimated amounts of deposit withdrawals during financial crises suggest that each branch should adopt a customer strategy appropriate for the variety of customers of that branch. The bank run model proposed in this research can also be applied to other marketing strategy planning, and has wide applicability.

## 1 Introduction[1]

The banking industry manages vast financial assets, directly provides various services to customers, and is one of the groups of companies which have the greatest impacts in market economies. Many financial crises have occurred in Asia, Europe, and America in recent years, with bank runs occurring each time. Not only banks which lack competitive strength, competitive banks also must consider responses to financial crises. This is because baseless rumors can also trigger a financial crisis. Particularly in Japan, deregulation of financial services has led to harsh competition in the banking industry. These kinds of changes are rapid, and in this harshly competitive environment, management to control risks in crisis situations [3] is a source of competitive advantage for bank continuity.

Many researches based on various approaches have been conducted on social and economic responses to financial crises. In macro level views, some researches investigated the impacts of the financial crises on markets after the government implementation of bank restructuring measures during financial crises in Thailand [8]

[1] Research of this paper was partly supported by the Grant-in-Aid for Science Research on Priority Areas (#19024048) and RCSS research fund by and the Ministry of Education, Science, Sports and Culture of Japan.

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 703–710, 2008.

and Korea [2]. These researches provided important suggestions on measures which the governments should take in the financial crises.

A typical research to analyze financial crises at a company level is on the relationship between corporate governance and stock valuation. Mitton [7], La Porta et al. [5], and Johnson et al. [4] investigated the differences that arise in stock valuation after financial crises, depending on the level of company disclosure, structure of ownership, level of diversification, and etc. More concretely, Anuchitworawong [1] analyzed the case of Thailand by investigating the relationship between owner behavior and stock prices after the financial crisis. This study clarified problems of undeveloped corporate governance in developing countries.

While these studies characterized the social and economic responses to the financial crises and suggested some effective measures to be taken at country and company levels, to our best knowledge, only few studies to characterize the customer behaviors in the financial crises and to suggest some counter measures in a bank branch level have reported. The objective of this research is to build a bank run model which highly accurately classifies whether or not individuals will withdraw all deposits from their accounts in banks for which they have obtained information about a bank fear. Particularly, we focus on the impacts of customer knowledge on deposit operations at the crises under the perspective of customer relationship management (CRM) of banks.

Because bank runs during financial crises are impacted by various factors of customers and their obtainable information, we can quantify those causal relationships by using individual level data. If such causality is understood, we can estimate bank runs by account holders during a financial crisis and provide important suggestions for customer management of banks against the account holder's response to a crisis. Furthermore, this model can also be applied to customer management in daily operations. To do this, we performed a questionnaire survey to collect data on the psychological responses of account holders to a bank fear. By using the model we built, we were able to obtain valuable information for banks to plan appropriate branch management strategies for financial crises.

The organization of this paper is as follows. We first build a model for bank runs in response to credit insecurities of banks. Next using this model, we estimate amounts of cash withdrawals during times of credit insecurity based on various scenarios. Finally, we present the implications of our proposed bank run model.

## 2 Bank Run Model

### 2.1 Data Used in Bank Run Model

In order to build the bank run model, data is required on individual attributes of bank account holders, and on various responses to credit information which cause bank runs, etc. In this paper we used questionnaire survey data on financial panics, which was collected in an internet survey. The questionnaire included basic attributes of respondents, such as gender, age, and education, and attributes related to bank run conduct, such recognition of deposit insurance, income level, amount of savings, and responses to insecurity regarding the banks with which transactions are done (withdraw deposits or not). The data collection period was from September 8 to 9, 2007. The survey method was a questionnaire survey via the internet. There were 1,500 samples.

## 2.2 Bank Run Modeling

We adopt as our target variable: Whether an account holder will withdraw his entire deposit or not, after receiving information on credit insecurity of his bank. We assume the sources of information on credit insecurities are TV, newspapers, magazines, internet, email, neighbors, and people at the workplace. This investigates whether, after receiving information on credit insecurities, the respondent is resolved to withdraw his entire bank deposit, or is resolved to not withdraw his entire deposit. Other actions are not considered (i.e. withdraw part of deposit).

**Table 1.** Attributes selected by *Ranker*

| Attribute | Avg. merit | Avg. rank |
|---|---:|---:|
| Form of employment | 0.015 | 2.5 |
| Marital status | 0.014 | 3.5 |
| Sex | 0.013 | 3.6 |
| Recognition of deposit insurance | 0.01 | 5.3 |
| Total deposits | 0.01 | 6 |
| Phone call frequency | 0.007 | 8.6 |
| Housing location | 0.007 | 8.8 |
| Trust in TV | 0.007 | 10.3 |
| Trust in Internet | 0.006 | 10.9 |
| Trust in neighbors | 0.006 | 10.9 |
| Annual income | 0.006 | 12.6 |
| Trust in workplace | 0.005 | 13.1 |
| Frequency of communication at the workplace | 0.005 | 13.2 |

The *Ranker* attribute selection algorithm [9] was used to select useful attributes from the questions mentioned above. Table 1 shows the attribute group extracted by *Ranker*. According to Table 1, it seems that form of employment is the most closely related to deposit withdrawal conduct in response to credit insecurities. For example, people are more likely to withdraw if they work in a family business, or are unemployed but seeking work. 13 attributes are selected, which also include marital status, sex, knowledge of deposit insurance, etc.

Next, we investigate the relationship in the bank run model between deposit withdrawal behavior in response to credit insecurities, and the explanatory variables such as form of employment. We build a model which estimates the probability of each consumer withdrawing deposits under certain conditions. The target variable is whether or not the deposit is withdrawn. This paper adopts a logit distribution (1) for the form of coefficient of the target variable.

$P_i$: Probability that Customer i will withdraw all deposits.
$B_m$: Attributes which explain deposit withdrawal behavior ($m$=1,2,…13).
$a_m$: Parameter of attribute $m$.

$$P_i = \frac{\exp(a_1B_1 + a_2B_2 + \cdots + a_mB_m)}{1 + \exp(a_1B_1 + a_2B_2 + \cdots + a_mB_m)} \,. \tag{1}$$

In order to obtain a superior prediction model, we applied machine learning techniques, and compared their accuracy with the logit model,. In this paper, we used a decision tree (C4.5), NaiveBayes, and a neural network (NN) as representative machine learning techniques for comparisons.

### 2.3 Resulting Estimations by the Bank Run Model

We use subject data to estimate the parameters sought for the logit model, and obtained statistically significant results. For the model's performance evaluation, we used *Overall Accuracy* [9,10], and the Matthews Correlation Coefficient (*C*) [6]. After classification, correctly classified positive and negative samples are $P_T$ and $N_T$ respectively. Incorrectly classified samples are $P_F$ and $N_F$. Overall Accuracy and C are defined as follows.

$$OverallAccuracy = \frac{P_T + N_T}{P_T + P_F + N_T + N_F} \,. \tag{2}$$

$$C = \frac{P_T \times N_T - P_F \times N_F}{\sqrt{(P_T + N_T)(N_T + P_F)(P_T + P_F)(N_T + N_F)}} \tag{3}$$

Fig. 1 shows indicators for evaluating the performance of the model obtained by 10 fold validation. As seen in Fig. 1, the prediction accuracy of the logit model is superior to that of the representative machine learning techniques C4.5, NaiveBayes, and NN. There were 0.4% of errors in estimation of the numbers of account holders who would participate in a bank run. It is relatively easy to interpret the logit model's estimates of parameters and its results. Therefore, in this paper we adopt the logit model for account holder behavior at times of credit insecurity, and to simulate outflows of bank held assets.

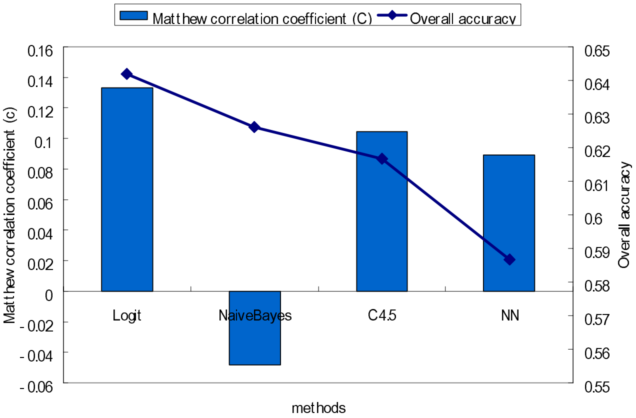


**Fig. 1.** Comparison of the prediction accuracy of each technique

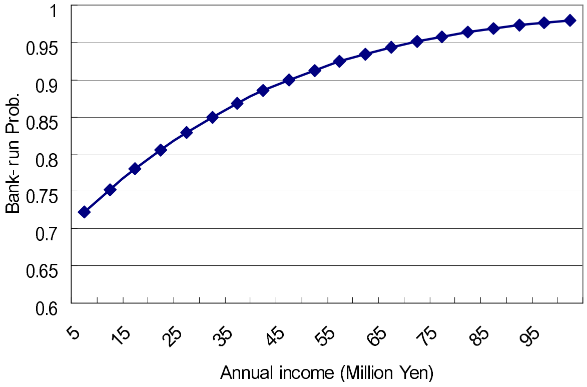


**Fig. 2.** Relationship between annual income and bank run behavior

Next, we use each value of the parameters obtained above with the logit model, and discuss the relationship between explanatory attributes and bank run behavior. Fig. 2 shows the relationship between annual income of account holders and their bank run behavior. As seen from the figure, one can conjecture that account holders who receive high annual income are more likely to participate in bank runs.

## 3 Estimation of Cash Reserves for Bank Runs

In order to handle bank runs in response to credit insecurities, each bank branch must prepare cash to be able to handle withdrawal demands of customers. By using the bank run model we built, it is possible to estimate the amount of cash reserve which each branch should hold. In this chapter, we present a process to estimate cash reserves in typical bank branches, based on the bank run model, and perform a simulation with that amount.

### 3.1 Basic Information on Bank Branches

In order to simulate cash reserve amounts, we investigate basic information on a typical bank branch. Banks do not necessarily have detailed information on all individual account holders. Of the limited information on individuals which banks have, the most accurate information is the existence of the account holders and their deposit amounts. In this paper, we assume that each individual has one account in the branch, and is able to withdraw the entire deposit amount at any time.

As a typical branch, we postulate Branch A which is in a dense residential area near a city center. Many houses are located within walking distance of the branch, and general salaried employees comprise a large percentage of the population. We assume this kind of branch has an average total of about 50,000 accounts, with an average deposit balance of 2.5 million yen. We also assume for the simulation that other explanatory attributes such as sex and marital status have the same distribution as among the questionnaire respondents. We use the above model to calculate the probability of bank run behavior of each customer, and total the average customer withdrawal amount for all branch customers, to estimate a total of 46.27 billion yen.

### 3.2 Simulation of a Variety of Branches

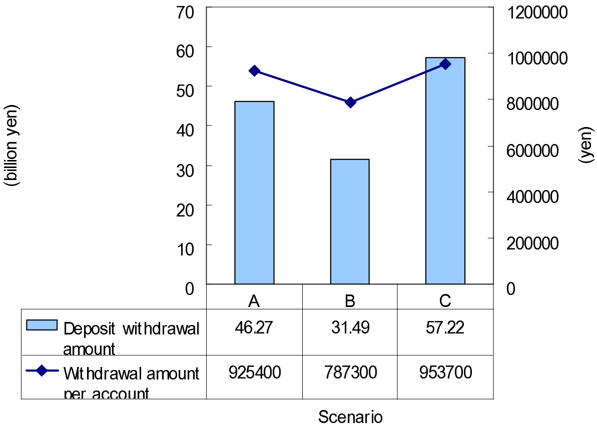
There are various differences in the environments where bank branches are located. Therefore, one can conjecture different distributions of account holders. Here, we clarify what differences in withdrawal behavior are brought about by these branch differences, and find whether this brings about resulting differences in estimated deposit withdrawal amounts. Since deposit insurance was implemented in Japan, customers move deposits between accounts to avoid risks, so average deposit balances are tending to become more uniform among all branches. Thus in this paper, we hypothesize 3 typical scenarios for a branch's number of accounts, location, and customers' forms of employment.

***Scenario A*:** Average urban branch. This is an average branch in a dense residential area as discussed in the previous section. Almost all account holders are salaried employees, the branch is located in a residential area, and the branch has about 50,000 accounts.

***Scenario B*:** Branch located in an area with many high class houses. The branch is located in an area with a high percentage of people with high incomes, with relatively few salaried employees. The branch as about 40,000 accounts, less than other branch types.

***Scenario C*:** Branch in an area with a high concentration of businesses. Self employed accounts comprise over 20% of all accounts, and there are about 60,000 accounts, more accounts than in other branch types.

Fig. 3 shows the total withdrawal amount and average withdrawal per account estimated in each scenario. As mentioned above, the bank run model is used to estimate



| | A | B | C |
|---|---|---|---|
| ▭ Deposit withdrawal amount | 46.27 | 31.49 | 57.22 |
| ◆ Withdrawal amount per account | 925400 | 787300 | 953700 |

**Fig. 3.** Differences between hypothesized branches in estimated deposit withdrawal amounts and withdrawal amounts per account

Scenario A total withdrawals of approximately 46.27 billion yen in response to credit insecurities. The estimates are 31.49 billion yen for Scenario B, and 57.22 billion yen for Scenario C. Even considering the different numbers of accounts, there are clear differences between scenarios regarding estimated total amount of deposit withdrawals. Scenario B has about 68% of the withdrawal amounts for Scenario A, while Scenario C has even 24% more than Scenario A. Regarding average withdrawal amounts per account, Scenario C has 953,700 yen, more than 20% greater than the 787,300 yen for Scenario B. In this way, we discovered differences in deposit withdrawal amounts in response to credit insecurities, corresponding to the environment such as the locations where bank branches are placed and customer characteristics.

## 4 Implications for Business

The bank run model proposed in this research has important practical suggestions for business. In the situation of the greatest crisis for banks, estimates of the deposit withdrawal behavior of customers provide important information for strategic planning to avoid serious risks. Use of this model enables estimates of the total cash which each bank branch should prepare in this kind of crisis. Also, responses of account holders differ by the location of branches, so this enables each branch to make a proper response corresponding to its situation.

The bank run model we presented and the framework for its construction are widely applicable. Various risks exist in the environments surrounding banks, from serious crises to minor problems. In this research, we focused on crises of serious credit insecurities, and performed a detailed analysis of account holder behavior. If one expands the framework which we presented, it is possible to build behavior models for individual customers in response to various risks. In short, this means that it is possible to estimate outflows and inflows of customer deposits in various situations. If bank branch managers use our model in daily operations, they can make current marketing activities more efficient. Our model is able to forecast individual behavior, so it is possible to apply in the development of highly trustworthy financial products.

## 5 Conclusion

In this paper, we used data from a questionnaire survey, modeled bank runs by account holders during times of credit insecurity, and clarified that responses differ according to the characteristics of customers. For example, we discovered that recognition of deposit insurance and customer form of employment are strongly related to bank run behavior. By using this model, we estimated the total amounts of deposit withdrawals from branches in times of credit insecurity. We found that differences in the location and main customer groups of a branch result in different estimated total deposit withdrawal amounts. We present a highly applicable model of bank runs in response to credit insecurities, and present a framework for quantitative understanding of consumer behavior in crisis situations.

However, several issues remain for the future.

1. Considering detailed simulations in varied scenarios.
2. To clarify consumer behavior in response to risks in situations other than financial crises, such as natural disasters or bank scandals.
3. To clarify differences in consumer behavior due to differences in sources of information on a bank fear.
4. To estimate the bias of the questionnaire survey which used the internet.

## References

1. Anuchitworawong, C.: Deposit insurance, corporate governance, and discretionary behavior: Evidence from Thai financial institutions, Centre for Economic Institute Working Paper # 2004-2015. Hitotsubashi University (2004)
2. Choe, H., Lee, B.S.: Korean bank governance reform after the Asian financial crisis. Pacific-Basin Finance Journal 11, 483–508 (2003)
3. Davies, G., Chun, R., Vinhas, R.S., Roper, S.: Corporate Reputation and Competitiveness. Routledge (2003)
4. Johnson, S., Boone, P., Breach, A., Friedman, E.: Corporate governance in the Asian financial crisis. Journal of Financial Economics 58, 141–186 (2000)
5. La Porta, R., Lopez-de-Silanes, F., Shleifer, A.: Corporate ownership around the world. Journal of Finance 54, 471–517 (1999)
6. Mattsatsinis, N., Siskos, Y.: Intelligent Support Systems for Marketing Decisions. Springer, Heidelberg (2002)
7. Mitton, T.: A cross-firm analysis of the impact of corporate governance on the East Asian financial crisis. Journal of Financial Economics 64, 215–241 (2002)
8. Pathan, S., Skully, M., Wickramanayake, J.: Reforms in Thai bank governance: The aftermath of the Asian financial crisis. International Review of Financial Analysis (in Press) (2008)
9. Witten, I.H., Frank, E.: Data Mining –Practical Machine Learning Tools and Techniques. Elsevier, Amsterdam (2005)
10. Yada, K., Ip, E., Katoh, N.: Is this brand ephemeral? A multivariate tree-based decision analysis of new product sustainability. Decision Support Systems 44(1), 223–234 (2007)

# Logic of Discovery and Knowledge: Decision Algorithm

Sergey Babenyshev and Vladimir Rybakov

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

**Abstract.** The logic of *Chance Discovery* (CD) as well as mathematical models for CD, by the very nature of the term *chance*, are hard to formalize, which poses challenging problems for mathematization of the area. It does not completely prevent us though from studying the *logical laws* which chance discovery and related notions should abide, especially in a carefully chosen and reasonably expressive mathematical formalism. The framework, the authors suggest[1] in this paper, is based on a well-developed area of modal logic, more precisely on Kripke-Hintikka semantics, with a notable distinction: unlike some other hybridization schemes, it leads to decidable logics, while still preserving high expressive power. We demonstrate our approach by an example of the *Logic of Discovery and Knowledge*, where a regular modal language is augmented with higher level operators intended to model some contrasting aspects of Chance Discovery: *uncertain necessity of discovery* and *local common knowledge* within contexts admitting branching time.

**Keywords:** chance discovery, modal logic, decidability, Kripke-Hintikka models, inference rules, rules in normal reduced form.

## 1 Introduction

In our paper we attempt to apply the techniques of non-classical logics to model various aspects of Chance Discovery (CD). Chance Discovery (cf. Ohsawa and McBurney [14], Abe and Ohsawa [1]) is a contemporary direction in Artificial Intelligence (AI) which analyzes important events with uncertain information, incomplete past data, so to say, *chance* events, where *a chance* is defined as some event which is significant for decision-making in a specified domain. Such events are typically rare and hard to identify. Chance Discovery is defined by Y.Ohsawa as “learning of or explaining a chance event”. The aim

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

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 711–718, 2008.

of CD as a discipline within AI is to determine methods for discovering the chance events.

The study of non-classical logics form a modern branch of mathematical logic, which has diverse application in Computer Science and Artificial Intelligence. Techniques of non-classical logic have been used to formalize various aspects of problems for computerized modeling of intelligent behavior. Significant part of this approach is based on modal-type logics, which were introduced at the beginning of the previous century. Major part of research was focused on reasoning about knowledge, time and computation (cf., for instance, Goldblatt [9], van Benthem [24]). Semantic tools for modal logic are often based on Kripke-Hintikka models and temporal algebras (cf. Thomason [23,10]). We aim to apply such technique to model various representations of Chance Discovery.

The notion of *chance*, by the nature of the term, is hard to describe and may be impossible to formalize in full at all, as contrasting meanings of words *chance* and *method* may suggest (see also M.Alai [2], M.Alai and G.Tarozzi [3]). However practical applications strongly require at least some formal tools of describing the situations of Chance Discovery. This pose a difficult task of formulating the logic of Chance Discovery, difficult but not impossible. One way to approach this problem is as follows. The major obstacle for recognizing a chance is incomplete nature of information at any given time moment. But for building a logic we can assume that we already have all information from the past and the future. The area of temporal and modal logics provides a wealth of useful and practical formal models, out of them one, Kripke-Hintikka semantics, has an advantage of the transparent epistemic and philosophical connections. So, based on a propositional modal logic defined semantically by a class of Kripke models, we formulate two cognitive operators available to a individual, one is just a *local common knowledge* operator and another is meant to represent *uncertain necessity of discovery* [22]. What makes these operators different from more traditional modal operators is that their formal definition in terms of underlying accessibility relation involves variables for *clusters* of the underlying frame and generally requires a second-order language.

To account for non-deterministic nature of reality we base our construction on modal logic S4. To represent incompleteness and uncertainty of the individual knowledge we assume that every time point is in fact a cluster of possible states of affairs. We check that chosen interpretations do not lead to "over expressiveness" of the language, by showing that logic LDK—Logic of Discovery and Knowledge, defined by all such models is decidable. In mathematical terms, we have found that LDK is decidable, and the satisfiability problem for it is solvable. The suggested method has already been proven to be powerful enough to help in a number of cases of epistemic logics [15,16,17,18,19,20,21,22]. Therefore it can be recommended as the method of first choice to study other possible formalizations of the related notions. We would like to emphasize that we do not develop any technique for extracting new information using CD, rather we give an algorithm to compute logical laws to which various formalizations of CD obey.

## 2 Notation, Preliminaries

The paper uses standard notation and known facts concerning modal, multimodal and temporal logics, and some familiarity with basic facts is assumed, though we give below all necessary definitions to follow the paper.

To formulate the logic LDK we proceed by introducing a semantic motivation for choice of its language and rules for computing truth values of formulas. The model of chance discovery, which we consider in this paper, is based on the Kripke-Hintikka frames $C_Y := \langle \bigcup_{i \in Y} C(i), R \rangle$, where $Y = \langle Y, \leqslant \rangle$ can be any partially ordered set, each $i \in Y$ is the time index for a finite set of possible states $C(i)$. In our formalism, any $C(i)$ is simply a set of elements (*worlds* or *states of affairs* in terms of Kripke-Hintikka semantics). A *branching-time flow (a non-deterministic computational process, possible evolutions of a system, a discovery search)* is modeled by the binary accessibility relation $R$ in $C_Y := \langle \bigcup_{i \in Y} C(i), R \rangle$, where for all elements $a$ and $b$ from $\bigcup_{i \in Y} C(i)$,

$$aRb \iff \exists i, j \in Y : i \leqslant j \; \& \; a \in C(i) \; \& \; b \in C(j).$$

Less formally, $R$ imitates the flow of time by connecting states, so, $aRb$ means that $a$ and $b$ are some states at the same time point or the state $b$ might be achieved after the time point where the state $a$ was present. How could we model the chance of discovery in these frames? First of all, we need to indicate possible events in models based on frames $C_Y$. For this we consider a set $P$ of facts, propositions, which may or may not happen, may be true or nor true in time flow. For each frame $C_Y$, we consider valuations $V$ of variables $P$, which are mappings of $P$ into the set of all subsets of the set $\bigcup_{i \in Y} C(i)$, so symbolically

$$\forall p \in P : V(p) \subseteq \bigcup_{i \in Y} C(i).$$

If, for an element $a \in \bigcup_{i \in Y} C(i)$, $a \in V(p)$ we say *the fact $p$ was discovered at the state $a$*. How to say that it is possible to discover a fact $p$ in the future research? The following operator was suggested (the operator $\Box$) in [22])

*it is necessary that the fact $\varphi$ may be discovered in any time point.*

We denote it $\mathcal{D}$ in this paper, to reserve the symbol $\Box$ for more traditional role of adjoint for $\Diamond$: $\Box := \neg\Diamond\neg$. Another operator we would like to augment our language with is $\mathcal{K}$ — the operator of *local common knowledge*. The meaning of $\mathcal{K}\varphi$ is *$\varphi$ is valid at any possible state of affairs at the current time point.*

The formal definition of formulas is as follows:

1. any propositional letter from $P$ is a well formed formula (wff),
2. if $\varphi$ and $\psi$ are wff's then $\varphi \wedge \psi$, $\varphi \vee \psi$, $\varphi \rightarrow \psi$ and $\neg\varphi$ are also wff's,
3. if $\varphi$ is a wff then $\Diamond\varphi$, $\mathcal{D}\varphi$ and $\mathcal{K}\varphi$ are also wffs.

Now we define rules for computing truth values of formulas in models $C_Y$ with valuations $V$ of propositions $P$. In the notation below, $(C_Y, a) \Vdash_V \varphi$ is meant to say that the formula $\varphi$ *is true at the state a in the model* $C_Y$ *w.r.t. valuation* $V$. The rules are as follows:

$$\forall p \in P,\ \forall a \in C_Y \quad (C_Y, a) \Vdash_V p \iff a \in V(p);$$
$$(C_Y, a) \Vdash_V \varphi \wedge \psi \iff [(C_Y, a) \Vdash_V \varphi] \text{ and } [(C_Y, a) \Vdash_V \psi];$$
$$(C_Y, a) \Vdash_V \varphi \vee \psi \iff [(C_Y, a) \Vdash_V \varphi] \text{ or } [(C_Y, a) \Vdash_V \psi];$$
$$(C_Y, a) \Vdash_V \varphi \to \psi \iff [(C_Y, a) \nVdash_V \varphi] \text{ or } [(C_Y, a) \Vdash_V \psi];$$
$$(C_Y, a) \Vdash_V \neg\varphi \iff (C_Y, a) \nVdash_V \varphi;$$
$$(C_Y, a) \Vdash_V \Diamond\varphi \iff \exists b \in C_Y\, [(aRb) \text{ and } (C_Y, b) \Vdash_V \varphi];$$
$$(C_Y, a) \Vdash_V \mathcal{D}\varphi \iff [a \in C(i) \implies [\forall j \geq i \exists b \in C(j) : (C_Y, b) \Vdash_V \varphi];$$
$$(C_Y, a) \Vdash_V \mathcal{K}\varphi \iff [a \in C(i)\ \&\ \forall b \in C(i) : (C_Y, b) \Vdash_V \varphi];$$

The truth evaluation for Boolean operations is quite standard, and the same holds for the operation *possible to discover*—$\Diamond$. Indeed, the statement $(C_Y, a) \Vdash_V \Diamond\varphi$ means that the fact (formula) $\varphi$ may be discovered — there is a future state for the current time point where the fact $\varphi$ is discovered (the formula $\varphi$ is true). Thus the operation $\mathcal{K}$ is a variant of the *week necessity* [22]. $\mathcal{D}\varphi$ does not say that $\varphi$ is always true, but it says that $\varphi$ is discoverable at some state of any time point (so, it says that there is always a chance to discover $\varphi$).

**Definition 1.** *The logic* LDK *is the set of all formulas which are true at any state of any frame* $C_Y$ *w.r.t. any valuation.*

What is immediately obvious is that the $\Diamond$-fragment of LDK is equal to $\Diamond$-fragment of the well known standard modal logic S4—it follows directly from the finite model property of S4.

We summarize the properties of $\mathcal{D}$ and $\mathcal{K}$ in the following lemmas

**Lemma 1.** *The following holds*

1. $\Box(p \to q) \to (\mathcal{D}p \to \mathcal{D}q) \in \text{LDK}$,
2. $\mathcal{D}p \to \mathcal{D}\mathcal{D}p \in \text{LDK}$,
3. $\varphi \in \text{LDK} \implies \mathcal{D}\varphi \in \text{LDK}$,
4. $\mathcal{D}\varphi \to \varphi \notin \text{LDK}$.

**Lemma 2.** *The following holds*

1. $\mathcal{K}(p \to q) \to (\mathcal{K}p \to \mathcal{K}q) \in \text{LDK}$,
2. $\mathcal{K}p \to \mathcal{K}\mathcal{K}p \in \text{LDK}$,
3. $\varphi \in \text{LDK} \implies \mathcal{K}\varphi \in \text{LDK}$,
4. $\mathcal{K}\varphi \to \varphi \in \text{LDK}$.
5. $\neg\mathcal{K}\neg\varphi \to \mathcal{K}\neg\mathcal{K}\neg\varphi \in \text{LDK}$.

So the $\mathcal{K}$-fragment of LDK coincides with modal logic S5.

Now, when the logic LDK has been defined, we would like to know which logical laws hold in it. For instance, whether two formulas $\varphi$ and $\psi$ are equivalent, whether a formula $\varphi$ is a theorem (a logical law) of LDK. A method for answering these questions is presented in the next section.

## 3 Results, Decidability

The question we will be dealing with in the sequel is how, for any given formula $\varphi$, to determine whether or not $\varphi$ is a theorem of LDK. If there is an algorithm for solving this problem then it is said that the logic is *decidable* and the algorithm provides a *decision procedure* for the logic. The logic LDK is a special modal logic, with additional modalities "aligned" with time, therefore we can use a previously developed technique [15,16,17,18,19,20,21,22] to tackle the problem. We remind now the definitions and notation. To avoid imminent numerous "proofs by induction" we will use a representation of formulas by rules in, so-called, *reduced normal form.* Recall that a (sequential) rule is an expression

$$r = \frac{\varphi_1(x_1,\ldots,x_n),\ldots,\varphi_m(x_1,\ldots,x_n)}{\psi(x_1,\ldots,x_n)},$$

where $\varphi_1,\ldots,\varphi_m,\psi$ are formulas built over some variables $x_1,\ldots,x_n$ and we write $\mathrm{Var}(r)=\{x_1,\ldots,x_n\}$. These variables $x_1,\ldots,x_n$ are called *variables of* $r$. Since we have conjunction in our language, we can consider only rules with one-formula premise.

A formula $\varphi$ is *valid in a frame* $C_Y$ (notation $C_Y \Vdash \varphi$) if, for any valuation $V$ of $\mathrm{Var}(\varphi)$ and for any element $a$ of $C_Y$, $(C_Y,a)\Vdash_V \varphi$. We write $C_Y \Vdash \varphi$ to indicate this fact.

**Definition 2.** *A rule* $r$ *is said to be* valid *in the Kripke model* $\langle C_Y, V\rangle$ *with the valuation* $V$ *(we will use notation* $C_Y \Vdash_V r$*) if*

$$\forall a : (C_Y,a)\Vdash_V \bigwedge_{1\le i\le m}\varphi_i \implies \forall a : (C_Y,a)\Vdash_V \psi.$$

*Otherwise we say* $r$ *is* refuted *in* $C_Y$*, or* refuted in $C_Y$ by $V$*, and write* $C_Y \nVdash_V r$.

A rule $r$ is *valid* in a frame $C_Y$ (notation $C_Y \Vdash r$) if, for any valuation $V$ of $\mathrm{Var}(r)$, $C_Y \Vdash_V r$. A rule $r$ is said to have the *reduced normal form* if $r=\varepsilon_r/x_1$ where

$$\varepsilon_r := \bigvee_{1\le j\le m}\bigwedge_{1\le i\le n}\left(x_i^{t(j,i,0)}\wedge(\mathcal{D}x_i)^{t(j,i,1)}\wedge(\mathcal{K}x_i)^{t(j,i,2)}\wedge(\Diamond x_i)^{t(j,i,3)}\right),$$

and $x_i$'s are certain letters (variables), $t(j,i,z)\in\{0,1\}$ and, for any formula $\alpha$ above, $\alpha^0 := \alpha$, $\alpha^1 := \neg\alpha$.

For any formula $\varphi$ we can convert it into the rule $x\to x/\varphi$ and then transform the latter into the reduced normal form.

**Definition 3.** *Given a rule* $r_{\mathrm{nf}}$ *in the reduced normal form,* $r_{\mathrm{nf}}$ *is said to be a normal reduced form for a rule* $r$ *iff, for any frame* $C_Y$*,* $C_Y \Vdash r \iff C_Y \Vdash r_{\mathrm{nf}}$*.*

Based on proofs of Lemma 3.1.3 and Theorem 3.1.11 from [16], by similar technique, we obtain

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

It is immediate to see that a formula $\varphi$ is valid in a frame $C_Y$ iff the rule $x \to x/\varphi$ is valid in $C_Y$, so from Theorem 1 we obtain

**Proposition 1.** *A formula* $\varphi$ *is a theorem of* LDK *iff the rule* $(x \to x/\varphi)_{nf}$ *is valid in any frame* $C_Y$*.*

**Lemma 3.** *A rule* $r_{\mathrm{nf}}$ *in the reduced normal form is refuted in a frame* $C_Y$ *w.r.t. a valuation* $V$ *if and only if* $r_{\mathrm{nf}}$ *is refuted in a frame* $C'_{Y'}$ *by a valuation* $V'$*, where*

1. *the size of any cluster* $C'(i)$ *in* $C'_{Y'}$ *is linear in the size of* $r_{\mathrm{nf}}$*;*
2. *the size of* $Y'$ *is less then* $2^m$*;*
3. *the size of the frame* $C'_{Y'}$ *is double exponentials in the size of* $r_{\mathrm{nf}}$*.*

Combining Theorem 1, Proposition 1 and Lemma 3 we derive

**Theorem 2.** *The logic* LDK *is decidable.*

The algorithm of verifying whether a formula is a theorem consists of checking the validity of the corresponding rule in the reduced normal form in Kripke frames of size effectively bounded by the size of the reduced normal form. The overall complexity also counts the price of reduction of rules to normal reduced forms, but this complexity is single exponential.

## 4 Conclusion

In the paper we suggest a modal framework suitable for formalizing some notions related to the field of Chance Discovery. It is based on well-understood Kripke-Hintikka semantics, but allows to introduce finitely many additional highly expressive operators. To illustrate our approach we consider the *Logic of Discovery and Knowledge*, where modal language is augmented with higher level operators intended to model some contrasting aspects of Chance Discovery: *uncertain necessity of discovery* and *local common knowledge* within contexts admitting branching time. In spite of high expressive power of additional operators, LDK preserves good mathematical properties of the original S4. In particular, we prove that the proposed interpretations are not "overly expressive" and do not lead to undecidability problems. In mathematical terms, we have found that LDK is decidable, and the satisfiability problem for it is solvable.

Another important and yet unsolved problem is finding explicit axiomatizations for similar constructions. Such an axiomatization would provide a desired clarification between concepts being modeled.

## References

1. Abe, A., Ohsawa, Y. (eds.): Readings in Chance Discovery. International Series on Advanced Intelligence (2005)
2. Alai M.: Univesity of Urbiono (manuscript, 2006), http://www.uniurb.it/Filosofia/isonomia/alai.pdf
3. Alai, M., Tarozzi, G. (eds.): Karl Popper Philosopher of Science. Rubbettino Editore, Soveria Mannelli (2006)
4. Arisha, K., Ozcan, F., Ross, R., Subrahmanian, V.S., Eiter, T., Kraus, S.: Impact: A platform for collaborating agents. IEEE Intelligent Systems 14(2), 64–72 (1999)
5. Avouris, N.M.: Co-operation knowledge-based systems for environmental decision-support. Knowledge-Based Systems 8(1), 39–53 (1995)
6. Bugarin, A.J., Barro, S.: Fuzzy reasoning supported by Petri nets. IEEE Transactions on Fuzzy Systems 2(2), 135–150 (1994)
7. Cortés, U., Sánchez-Marré, M., Ceccaroni, L.: Artificial intelligence and environmental decision support systems. Applied Intelligence 13(1), 77–91 (2000)
8. Dubois, Prade, H.: The three semantics of fuzzy sets. Fuzzy Sets and Systems 90(2), 141–150 (1997)
9. Goldblatt, R.: Logics of Time and Computation. CSLI Lecture Notes 7 (1992)
10. Goldblatt, R.: Mathematical Modal Logic: A View of its Evolution. J. Applied Logic 1(5–6), 309–392 (2003)
11. Harmelem, F., Horrocks, I.: The semantic web and its languages faqs on oil: The ontology inference layer. IEEE Intelligent Systems 15(6), 69–72 (2000)
12. Hintikka, J., Vandamme, F.: Logic of Discovery and Logic of Discourse. Springer, Heidelberg (1986)
13. Hendler, J.: Agents and the semantic web. IEEE Intelligent Systems 16(2), 30–37 (2001)
14. Ohsawa, Y., McBurney, P. (eds.): Chance Discovery (Advanced Information Processing). Springer, Heidelberg (2003)
15. Rybakov, V.V.: A Criterion for Admissibility of Rules in the Modal System S4 and the Intuitionistic Logic. Algebra and Logic 23(5), 369–384 (1984) (Engl. Translation)
16. Rybakov, V.V.: Admissible Logical Inference Rules. Studies in Logic and the Foundations of Mathematics, vol. 136. Elsevier Sci. Publ. North-Holland, Amsterdam (1997)
17. Rybakov, V.V.: Construction of an Explicit Basis for Rules Admissible in Modal System S4. Mathematical Logic Quarterly 47(4), 441–451 (2001)
18. Rybakov, V.V.: Logical Consecutions in Intransitive Temporal Linear Logic of Finite Intervals. Journal of Logic Computation 15(5), 633–657 (2005)
19. Rybakov, V.V.: Logical Consecutions in Discrete Linear Temporal Logic. Journal of Symbolic Logic 70(4), 1137–1149 (2005)
20. Rybakov, V.V.: Linear Temporal Logic with Until and Before on Integer Numbers, Deciding Algorithms. In: Grigoriev, D., Harrison, J., Hirsch, E.A. (eds.) CSR 2006. LNCS, vol. 3967, pp. 322–333. Springer, Heidelberg (2006)
21. Rybakov, V.V.: Branching Time Logic $\mathcal{PTL}_\alpha$ with Operations Until and Since Based on Bundles of Integer Numbers, Logical Consecutions, Deciding Algorithms (submitted, 2006)

22. Rybakov, V.V.: Logic of Discovery in Uncertain Situations—Deciding Algorithms. In: Apolloni, B., Howlett, R.J., Jain, L. (eds.) KES 2007, Part II. LNCS (LNAI), vol. 4693, pp. 950–958. Springer, Heidelberg (2007)
23. Thomason, S.K.: Semantic Analysis of Tense Logic. Journal of Symbolic Logic 37(1) (1972)
24. van Benthem, J.: The Logic of Time. Reidel, Dordrecht, Synthese Library, Vol. 156 (1983)

# Chances, Affordances, Niche Construction

Lorenzo Magnani

University of Pavia, Department of Philosophy and Computational Philosophy Laboratory, Pavia, Italy, and Sun Yat-sen University, Department of Philosophy, Guangzhou (Canton), P.R. China
lmagnani@unipv.it

**Abstract.** As a matter of fact, humans continuously delegate and distribute cognitive functions to the environment to lessen their limits. They build models, representations, and other various mediating structures, that are considered to aid thought. In doing these, humans are engaged in a process of cognitive niche construction. In this sense, I argue that a cognitive niche emerges from a network of continuous interplays between individuals and the environment, in which people alter and modify the environment by mimetically externalizing fleeting thoughts, private ideas, etc., into external supports. Through mimetic activities humans create *external semiotic anchors* that are the result of a process in which concepts, ideas, and thoughts are projected onto external structures. Once concepts and thoughts are externalized and projected, new *chances* and ways of inferring come up from the blend. For cognitive *niche construction* may also contribute to make available a great portion of knowledge that otherwise would remain simply unexpressed or unreachable.

## 1 Introduction

As a matter of fact, humans continuously delegate and distribute cognitive functions to the environment to lessen their limits. They build models, representations, and other various mediating structures, that are thought to be good to think. The aim of this paper is to shed light on these *designing* activities. In the first part of the paper I will argue that these designing activities are closely related to the process of niche construction. I will point out that in building various mediating structures, such as models or representations, humans alter the environment and thus create *cognitive niches*.

In dealing with the exploitation of cognitive resources and chances embedded in the environment, the notion of *affordance*, originally proposed by Gibson [1] to illustrate the hybrid character of visual perception, can be extremely relevant. The analysis of the concept of affordance also provides an alternative account about the role of external – also artifactual – objects and devices. Artifactual cognitive objects and devices extend, modify, or substitute "natural" affordances actively providing humans and many animals with new opportunities for action [2]. In order to solve various controversies on the concept of affordance, I will take advantage of some useful insights that come from the study on *abduction* [3]. Abduction may fruitfully describe all those human and animal hypothetical inferences that are operated through actions which consist in smart manipulations to both detect new affordances and to create manufactured external objects that offer new affordances.

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 719–726, 2008.

# 2 Humans as Chance Seekers

## 2.1 Incomplete Information and Human Cognition

Humans usually make decisions and solve problems relying on incomplete information [4]. Having incomplete information means that 1) our deliberations and decisions are never *the best* possible answer, but they are at least *satisficing*; 2) our conclusions are always *withdrawable* (i.e. questionable, or never final). That is, once we get more information about a certain situation we can always revise our previous decisions and think of alternative pathways that we could not "see" before; 3) a great part of our job is devoted to elaborating conjectures or hypotheses in order to obtain more adequate information. Making conjectures is essentially an *act* that in most cases consist in manipulating our problem, and the representation we have of it, so that we may eventually acquire/create more "valuable" knowledge resources. Conjectures can be either the fruit of an abductive selection in a set of pre-stored hypotheses or the creation of new ones, like in scientific discovery. To make conjectures humans often need more evidence/data: in many cases this further cognitive action is the only way to simply make possible (or at least enhance) a thought to "hypotheses" which are hard to successfully perform. In this sense, humans can be considered *chance seekers*, because they are continuously engaged in a process of building up and then extracting latent possibilities to uncover new valuable information and knowledge.

The idea I will try to deepen in the course of this paper is the following: as chance seekers, humans are *ecological engineers*. That is: humans like other creatures do not simply live their environment, but they actively shape and change it looking for suitable chances. In doing so, they construct *cognitive niches* [5,6,7] through which the offerings provided by the environment in terms of cognitive possibilities are appropriately selected and/or manufactured to enhance their fitness as chance seekers. Hence, this ecological approach aims at understanding cognitive systems in terms of their *environmental situatedness* [8,9]. Within this framework, chances are that "information" which is not stored internally in memory or already available in an external reserve but that has to be "extracted" and then *picked up* upon occasion.

As I will show in the following two sections, "the active interrogation of the environment" is also at the root of the evolution of our organism and its cognitive system. I will also describe this ecological activity by using the notion of abduction [3] and its semiotic dimension. (I have already treated the relationships between abduction and chance in [8,10]).

## 2.2 Cognitive Niche Construction and Human Cognition as a Chance-Seeker System

It is well-known that one of the main forces that shape the process of adaptation is natural selection. That is, the evolution of organisms can be viewed as the result of a selective pressure that renders them well–suited to their environments. Adaptation is therefore considered as a sort of *top-down process* that goes from the environment to the living creature [11]. In contrast to that, a small fraction of evolutionary biologists have recently tried to provide an alternative theoretical framework by emphasizing the role of niche construction [12] [13].

According to this view, the environment is a sort of "global market" that provides living creatures with unlimited possibilities. Indeed, not all the possibilities that the environment offers can be exploited by the human and non-human animals that act on it. For instance, the environment provides organisms with water to swim in, air to fly in, flat surfaces to walk on, and so on. However, no creatures are fully able to take advantage of all of them. Therefore, all organisms try to modify their surroundings in order to better exploit those elements that suit them and eliminate or mitigate the effect of the negative ones. This process of *environmental selection* [14] allows living creatures to build and shape the so-called "ecological niches". An ecological niche can be defined as a "setting of environmental features that are suitable for an animal" [1]. It differs from the notion of habitat in the sense that the niche describes *how* an organism lives its environment, whereas habitat simply describes *where* an organism lives. In any ecological niche, the selective pressure of the *local* environment is drastically modified by organisms in order to lessen the negative impacts of all those elements which they are not suited to. Indeed, this does not mean that natural selection is somehow halted.

My contention is that the notion of niche construction can be also usefully applied to human cognition. More precisely, I claim that cognitive niche construction can be considered as one of the most distinctive traits of human cognition, i.e. humans construct "cognitive niches".[1] It emerges from a network of continuous interplay between individuals and the environment, in which they more or less tacitly manipulate what is occurring outside at the level of the various structures of the environment in a way that is suited to them. Accordingly, we may argue that the creation of cognitive niches is *the* way cognition evolves, and humans can be considered as ecological cognitive engineers.

Recent studies on distributed cognition seem to support my claim [16,17,18,19]. According to this approach, cognitive activities like, for instance, problem solving or decision-making, cannot only be regarded as internal processes that occur within the isolated brain. Through the process of niche creation humans extend their minds into the material world, exploiting various external resources. For "external resources" I mean everything that is not inside the human brain, and that could be of some help in the process of deciding, thinking about, or using something. Therefore, external resources can be artifacts, tools, objects, and so on. They are property of individuals in so far as those individuals are embedded in given settings or environments [20]. Something important must still be added, and it deals with the notion of representation: the traditional notion of representation as a kind of abstract mental structure is old-fashioned and misleading [21]. If some cognitive performances can be viewed as the result of a smart interplay between humans and the environment, the representation of a problem is partly internal but it also depends on the smart interplay between the individual and the environment.

As I have already said, an alternative definition of the ecological niche that I find appealing in treating our problem has been provided by Gibson [1]: he pointed out that a

[1] If we also recognize in animals, like many ethologists do, a kind of nonlinguistic thinking activity basically model-based (i.e. devoid of the cognitive functions provided by human language), their ecological niches can be called "cognitive", when for example complicate animal artifacts like landmarks of caches for food are fruit of "flexible" and learned thinking activities which indeed cannot be entirely connected with innate endowments [15].

niche can be seen as a set of *affordances*. My contention is that the notion of affordance may help provide sound answers to the various questions that come up with the problem of ecological niches. The notion of affordance is fundamental for two reasons. First of all, it defines the nature of the relationship between an agent and its environment, and the mutuality between them. Second, this notion may provide a general framework to illustrate humans as chance seekers.

## 3 Affordances as Eco-cognitive Interactional Chances

As I have illustrated in the first part of this paper, humans and some animals manipulate and distribute cognitive meanings after having delegated them to suitable environmental supports. The activity of cognitive niche construction reveals something important about the human and animal cognitive system. As already mentioned, human cognition can be better understood in terms of its environmental situatedness. This means humans do not retain in their memory an explicit and complete representation of the environment and its variables, but they actively manipulate it by picking up information and resources upon occasion. Information and resources are not only given, but they are actively sought and even manufactured. In this sense, I consider human cognition as a chance-seeker system. In my terminology, chances are not simply information, but they are "affordances", namely, *environmental anchors* that allow us to better exploit external resources.

One of the most disturbing problems with the notion of affordance is that any examples provide different, and sometimes ambiguous insights on it. This fact makes very hard to give a conceptual account of it. That is to say, when making examples everybody grasps the meaning, but as soon as one tries to conceptualize it the clear idea one got from it immediately disappears. Therefore, I hope to go back to examples from abstraction without loosing the intuitive simplicity that such examples provide to the intuitive notion.

Gibson defines "affordance" as what the environment offers, provides, or furnishes. For instance, a chair affords an opportunity for sitting, air breathing, water swimming, stairs climbing, and so on. Gibson did not only provide clear examples, but also a list of definitions that may contribute to generating possible misunderstanding: 1) affordances are opportunities for action; 2) affordances are the values and meaning of things which can be directly perceived; 3) affordances are ecological facts; 4) affordances imply the mutuality of perceiver and environment.

I contend that the Gibsonian ecological perspective originally achieves two important results. First of all, human and animal agencies are somehow hybrid, in the sense that they strongly rely on the environment and on what it offers. Secondly, Gibson provides a general framework about how organisms directly perceive objects and their affordances, as behavioral and cognitive chances. His hypothesis is highly stimulating: "[...] the perceiving of an affordance is not a process of perceiving a value-free physical object [...] it is a process of perceiving a value-rich ecological object", and then, "physics may be value free, but ecology is not" [1, p. 140]. These two issues are related, although some authors seem to have disregarded their complementary nature.

We can find that a very important aspect that is also disregarded in literature is the dynamic one, related to designing affordances with respect to their evolutionary framework: human and non-human animals can “modify” or “create” affordances by manipulating their cognitive niches. Moreover, it is obvious to note that human, biological bodies themselves evolve: and so we can guess that even the more basic and wired perceptive affordances available to our ancestors were very different from the present ones. Of course different affordances can be detected in children, and in the whole realm of animals.

Moreover, organisms need to become *attuned* to the relevant offered features and much of the cognitive tools built to reach this target are a result of evolution and of merely wired and embodied perceptual capacities like imagistic, empathetic, trial and error, and analogical devices. These capabilities, that in our epistemological perspective have to be considered “cognitive” even if instinctual, can be seen as devices of organisms that provide potential implicit abductive powers: they can provide an overall appraisal of the situation at hand and so orient action, and can be seen as forms of pseudo-explanation of what is occurring over there, as emerging in that material contact with the environment granted in the perceptual interplay. It is through this embodied process that affordances can arise as chances both in wild and artificially modified niches. Indeed, humans and animals make available – create – new affordances through the manipulation of the environment and the construction of artifacts; moreover these artifacts are often fruit of high-level – not merely reflex-based, instinctual – cognitive plastic endowments.

### 3.1 Affordances and Abduction

Let us consider the case of a chair that affords sitting. Now, my point is that we should distinguish between two cases: in the first one, the cues we come up with (flatness, robustness, rigidity) are *highly diagnostic* chances to know whether or not we can sit down on it, whereas in the second case we eventually decide to sit down, but we do not have any precise clue about. How many things are there that are flat, but one cannot sit down on? A nail head is flat, but it is not useful for sitting. This example introduces two important elements: firstly, finding/constructing affordances deals with a (semiotic) inferential activity; secondly, it distinguishes between an affordance and the information that specify it that only arise in the *eco-cognitive interaction* between environment and organisms. I maintain that the notion of abduction can clarify this puzzling problem.

The term “highly diagnostic” explicitly refers to the abductive framework. Abduction is a process of *inferring* certain facts and/or laws and hypotheses that render some sentences plausible, that *explain* or *discover* some (eventually new) phenomenon or observation. The distinction between theoretical and manipulative abduction extends the application of that concept beyond a sentential dimension. From Peirce’s philosophical point of view, all thinking is in signs, and signs can be icons, indices or symbols. Moreover, all inference is a form of sign activity, where the word sign includes “feeling, image, conception, and other representation” [22, 5.283], and, in Kantian words, all synthetic forms of cognition. That is, a considerable part of the thinking activity is “model-based”. Of course model-based reasoning acquires its peculiar creative relevance when embedded in abductive processes, so that we can individuate a *model-based abduction*, a considerable part of the “performances that involve sign activities are abductions” (*cit.*). [22, 5.283]. In the case of diagnostic reasoning, a physician detects various symptoms (that are signs or clues), for instance, cough, chest pain, and fever, *then* he/she may infer that it is a case of pneumonia.

The original Gibsonian notion of affordance deals with those situations in which the "perceptual" signs and clues we can detect prompt or suggest a certain action rather than others.[2] They are already available and belong to the normality of the adaptation of an organism to a given ecological niche. Nevertheless, if we acknowledge that environments and organisms' instinctual and cognitive plastic endowments change, we may argue that affordances can be related to the variable (degree of) *abductivity* of a configuration of signs: *a chair affords sitting* in the sense that the action of sitting is a result of a sign activity in which we perceive some physical properties (flatness, rigidity, etc.), and therefore we can ordinarily "infer" (in Peircean sense) that a possible way to cope with a chair is sitting on it. So to say, in most cases it is a spontaneous abduction to find affordances because this chance is already present in the perceptual and cognitive endowments of human and non-human animals.

I maintain that describing affordances that way may clarify some puzzling themes proposed by Gibson, especially the claim concerning the fact that we directly perceive affordances and that the value and meaning of a thing is clear on first glance. As I have just said, organisms have at their disposal a standard endowment of affordances (for instance through their wired sensory system, which is the only cognitive system "available" in the case of simple organisms), but at the same time they can extend and modify the range of what can afford them through the appropriate cognitive abductive skills (more or less sophisticated). As maintained by several authors (for example cf. [23,3]), what we see is a result of an embodied cognitive abductive process. For example, people are adept at imposing order on various, even ambiguous, stimuli [3, p. 107]. Roughly speaking, we may say that what we *see* is what our visual apparatus can, so to say, "explain". It is worth noting that this process happens almost simultaneously without any further mediation. Perceiving affordances has something in common with it. Visual perception is indeed a more automatic and "instinctual" activity, I have already said that Peirce claimed to be essentially abductive, even if not propositional. Indeed he considers inferential any cognitive activity whatever, not only conscious abstract thought: he also includes perceptual knowledge and subconscious cognitive activity.

We also have to remember that environments change and so the perceptive capacities when enriched through new or higher-level cognitive skills, which go beyond the ones granted by the merely instinctual levels. This dynamics explains the fact that if affordances are usually stabilized this does not mean they cannot be modified and changed and that new ones can be formed. First of all, affordances appear durable in human and animal behavior, like a kind of habit, as Peirce would say [22, 2.170]. For instance, that a chair affords sitting is a fair example of what I am talking about. This deals with

[2] In the original Gibsonian view the notion of affordance is strictly referred to proximal and immediate perceptual chances, which do not involve higher cognitive functions such as memory and schemata, but which are merely "picked up" by a stationary or moving observer. In this paper I maintain that perceiving affordances also involves evolutionary changes and the role of more sophisticated and plastic cognitive capacities. In an analogous perspective Norman [2] and Zhang and Patel [19] expand the notion into a framework for design and for distributed cognition.

what we may call *stabilized affordances*. That is, affordances that we have experienced as highly successful. Once evolutionarily formed, or created/discovered through cognition, they are stored in embodied or explicit cognitive libraries and retrieved upon occasion. Not only can they be a suitable source of new chances, through analogy. We may have very different objects that equally afford sitting. For instance, a chair has four legs, a back, and it also stands on its own. The affordances exhibited by a traditional chair may be an analogical source and transferred to different new artifacts that present the affordance of a chair for sitting down (and that to some extent can still be described as a chair). Consider, for instance, the variety of objects that afford sitting without having four legs or even a back. Let us consider a stool: it does not have even a back or, in some cases, it has only one leg or just a pedestal, but it affords sitting as well as a chair.

Second, affordances are also subjected to changes and modifications. Some of them can be discarded, because new artifacts are invented with more powerful ones. Consider, for instance, the case of blackboards. Progressively, teachers and instructors have partly replaced them with new affordances brought about by various tools, for example, slide presentations. In some cases, the affordances of blackboards have been totally re-directed or re-used to more specific purposes. For instance, one may say that a logical theorem is still easier to be explained and understood by using a blackboard, because of its affordances that give a temporal, sequential, and at the same time global perceptual depiction to the matter. In this perspective we acknowledge that artifacts like computational programs are tools for thoughts as is language: tools for exploring, expanding, and manipulating our own minds. The *information artifacts* [24] or *cognitive artifacts* [25] which represent the external multiple tools - communication, context shifting, computational devices expressly constructed to the aim of creating opportunities and risks, like KeyGraph, etc. - recently analyzed by the researchers in the field of chance discovery [26], are artifacts as *chance seeking prostheses* offered to humans to enhance their cognitive capabilities.

## 4 Conclusion

In this paper I have argued that cognitive niche construction is one of the most distinctive traits of human cognition. Humans and many other organisms continuously manipulate the environment in order to exploit its offerings. In doing this, they are engaged in a process of altering or even creating external structures to lessen and overcome their limits. New ways of coping with the environment, through both evolution and cultural evolution, namely affordances, are thus created. In dealing with cognitive niche construction, I have argued that abduction can fruitfully shed light on various aspects of building affordances.

## References

1. Gibson, J.J.: The Ecological Approach to Visual Perception. Houghton Mifflin, Boston (1979)
2. Norman, D.: The Design of Everyday Things. Addison Wesley, New York (1988)

3. Magnani, L.: Abduction, Reason, and Science. Processes of Discovery and Explanation. Kluwer Academic/Plenum Publishers, New York (2001)
4. Simon, H.: A behavioral model of rational choice. The Quarterly Journal of Economics 69, 99–118 (1955)
5. Tooby, J., DeVore, I.: The reconstruction of hominid behavioral evolution through strategic modeling. In: Kinzey, W.G. (ed.) Primate Models of Hominid Behavior, pp. 183–237. Suny Press, Albany (1987)
6. Pinker, S.: How the Mind Works. Norton, New York (1997)
7. Pinker, S.: Language as an adaptation to the cognitive niche. In: Christiansen, M.H., Kirby, S. (eds.) Language Evolution. Oxford University Press, Oxford (2003)
8. Magnani, L.: Chance discovery and the disembodiment of mind. In: Oehlmann, R., Abe, A., Ohsawa, Y. (eds.) Proceedings of the Workshop on Chance Discovery: from Data Interaction to Scenario Creation, International Conference on Machine Learning (ICML 2005), pp. 53–59 (2005)
9. Clancey, W.J.: Situated Cognition: on Human Knowledge and Computer Representations. Cambridge University Press, Cambridge (1997)
10. Magnani, L.: Prefiguring ethical chances: the role of moral mediators. In: Oshawa, Y., Tsumoto, S. (eds.) Chance Discoveries in Real World Decision Making: Data-based Interaction of Human and Artificial Intelligence, pp. 205–229. Springer, Berlin (2006)
11. Godfrey-Smith, P.: Complexity and the Function of Mind in Nature. Cambridge University Press, Cambridge (1998)
12. Laland, K., Odling-Smee, J., Feldman, M.W.: Niche construction, biological evolution and cultural change. Behavioral and Brain Sciences 23(1), 131–175 (2000)
13. Laland, K.N., Odling-Smee, F.J., Feldman, M.W.: Cultural niche construction and human evolution. Journal of Evolutionary Biology 14, 22–33 (2001)
14. Odling-Smee, J.J.: The Role of Behavior in Evolution. Cambridge University Press, Cambridge (1988)
15. Magnani, L.: Animal abduction. From mindless organisms to artifactual mediators. In: Magnani, L., Li, P. (eds.) Model-Based Reasoning in Science, Technology, and Medicine. Springer, Berlin (2007) (forthcoming)
16. Hutchins, E.: Cognition in the Wild. The MIT Press, Cambridge (1995)
17. Clark, A., Chalmers, D.J.: The extended mind. Analysis 58, 10–23 (1998)
18. Wilson, R.A.: Boundaries of the Mind. Cambridge University Press, Cambridge (2004)
19. Zhang, J., Patel, V.L.: Distributed cognition, representation, and affordance. Cognition & Pragmatics 2(14), 333–341
20. Wilson, R.A.: Collective memory, group minds, and the extended mind thesis. Cognitive Process 6(4), 429–457 (2005)
21. Gatti, A., Magnani, L.: On the representational role of the environment and on the cognitive nature of manipulations. In: Magnani, L., Dossena, R. (eds.) Computing, Philosophy and Cognition, pp. 227–242. College Publications, London (2005)
22. Peirce, C.S.: Collected Papers of Charles Sanders Peirce. Harvard University Press, Cambridge (1931-1958); vols. 1-6, Hartshorne, C., Weiss, P. (eds.); vols. 7-8, Burks, A.W. (ed.)
23. Hoffman, D.D.: Visual Intelligence: How We Create What We See. Norton, New York (1998)
24. Amitani, S., Hori, K.: A method and a system for supporting the process of chance discovery. In: Abe, A., Oehlmann, R. (eds.) The First European Workshop on Chance Discovery, Valencia, pp. 62–71 (2004)
25. Shibata, H., Hori, K.: Towards and integrated environment for writing. In: Abe, A., Oehlmann, R. (eds.) The First European Workshop on Chance Discovery, Valencia, pp. 232–241 (2004)
26. Oshawa, Y., McBurney, P. (eds.): Chance Discovery. Springer, Berlin (2003)

# Free Association Versus Recognition: Sensitivity for Chance Discovery in Cross-Cultural Color Design

Gyoung Soon Choi[1], Ruediger Oehlmann[2], and David Cottington[1]

[1] Kingston University London, Faculty of Art, Design and Architecture,
Kingston-upon-Thames, KT1 2QJ, UK
gsoonchoi@gmail.com
[2] Kingston University London, Faculty of Computing, Information Systems and Mathematics,
Kingston-upon-Thames, KT1 2EE, UK
R.Oehlmann@Kingston.ac.uk

**Abstract.** Chance Discovery focuses on opportunities and risks for future decision making. This makes it suitable for cross-cultural color design. An on-going research program aims at developing a new method of deriving opportunities and risks for design frameworks from the semantic and context analysis of the culturally constrained use of color. As part of this program, this paper compares two experiments involving two groups of Korean and English participants. In the first qualitative experiment participants were presented with preferred English and Korean colors and were asked to describe the meaning of the colors in a free association format. In the second quantitative experiment two different groups of English and Korean participants were presented with the same colors combined with selected explanations from the first experiment. The participants had then to identify their preference on a Likert scale from 1 to 7. Although the differences between colors ($p < 0.001$) and between meanings ($p < 0.001$) were significant, the differences between Korean and English participants were not significant ($p < 0.077$). The difference in the experimental results suggests that both methods have different sensitivities for the discovery of new cross-cultural chances. Implications for e-commerce Web design will be discussed.

**Keywords:** Discovering Color Semantics as a Chance, Free association, Recognition task, Cross-cultural meshing.

## 1 Introduction

Chance Discovery with its focus on the identification of opportunities and risks is of considerable use for designers. In particular in cross-cultural design, a strategy which may be suitable for one culture may bear risks in another culture. This applies also to the design of e-commerce Web sites. For instance, Hu et al. [1] have shown that consumers in China and Japan perceive Web sites differently from consumers in the UK. A color that draws attention to a product in Japan, may not have the same effect in the UK and *vice versa*. This type of result is highly relevant for the designers of e-commerce Web site, who wish to sell products across cultural divides. The designer would need to consider the cultural differences in various target markets.

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 727–733, 2008.

The study in this paper is part of an ongoing research program to develop an interactive design framework that will guide designers in the use of color [2]. It will aim at the cross-cultural dimension of design decisions, which has been neglected in current frameworks.

In order to determine cross-cultural differences, comparative research has to address the understanding of human psychological phenomena in different cultures [3]. Cross-cultural psychology is the empirical study and research based on different experiences in different cultural groups [4]. Cross-cultural differences and similarities have been investigated in controlled experiments. Comparing cultures requires the identification of causal relationships between independent and dependent variables from such experiments. The experiments need to be controlled for confounding variables to insure that the dependent variables depend only on the independent variables and not on any other effects. Only in this way it is possible to identify causal relations between variables that may not be apparent to the naked eye [5].

The color experience of consumers can be captured in an interactive framework, if such a framework accounts for semantic and contextual differences between cultures. Choi et al. [6] have investigated the meaning that Korean and English participants assign to a given color. This was essentially an experiment based on free associations. Participants associated the presented color with the first meaning that came to mind. The findings suggested that Korean participants form more associations related to emotions and nature, while English participants form more associations with artificial concept categories, such as man-made objects. This means that there was a clear cross-cultural distinction between the meanings associated by Korean and English participants.

The current study, replaced the free association task with a recognition task that asked participants in both cultures to make sense of presented color/meaning combinations by providing ratings indicating the appropriateness of a combination. It was therefore predicted that a meaning generated by Korean participants in the previous experiment will attract a high rating (ratings from 5 to 7) from Korean participants and a low rating (ratings from 1 to 3) from English participants. Equally a meaning generated by English participants in the previous experiment, will attract a high rating by English participants and a low rating by Korean participants.

The remainder of this paper will summarize the previous experiment in Section 2, followed by a description of the current experiment in Section 3. The discussion in Section 4 will consider the results and highlight the differences between the two experiments for general approaches in cross-cultural chance discovery.

## 2 A Previous Experiment on Cross-Cultural Color Semantics

The main objective of this experiment was to test the prediction that Korean participants form more associations related to emotions and nature, while English participants form more associations with artificial concept categories, such as man-made objects. In a previous experiment, 5 colors that were the most preferred by Korean participants and another 5 colors that were the most preferred by English participants have been determined [2]. These 10 colors were presented on a computer screen to 12 Korean and 12 English participants. Both groups were asked to describe in written form the first ideas and feelings that come to their mind when they look at a given color.

Responses could be classified in the categories warm emotion, cold emotion, spiritual, physical/material, nature, artificial, objects and others. The results confirmed the hypothesis. Qualitative differences of the colors ‘red’, ‘navy blue’, ‘dark green’, ‘bright green’, ‘white’, ’yellow’ were found in relation to cross-cultural differences between English and Korean participants. For example, the semantic descriptors of the color ‘red’ used by Korean participants were for instance ‘hot summer day’, ‘well matured Sharon fruits’, ‘weather between spring and summer’ and ‘sun’, which were categorized as warm emotion and nature. In contrast, English participants used descriptors such as ‘poppy’s field’, ‘plastic’, ‘football team’ and ‘modern design’, which are a wide range of artificial concepts and objects. The details of this previous experiment can be found in [6].

## 3 Current Experiment

The main objective of the current experiment was to investigate the preferred color/meaning combinations in the English and Korean cultures.

### 3.1 Method

#### 3.1.1 Design

The experiment was based on a mixed design with the culture (Korean or English) as unrelated and color selection, cultural origin of the meaning and the actual meaning as related variables. 5 the most preferred Korean colors and another 5 the most preferred English colors together with the most informative meanings were presented to two groups of 50 participants each. The first group consisted of English participants and the second group of Korean participants.

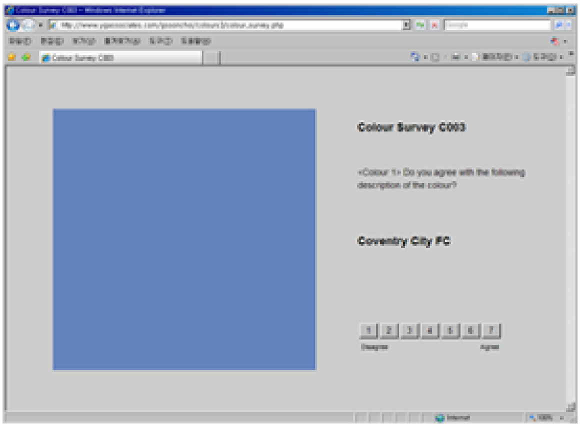
#### 3.1.2 Participants

Participants in the Korean group were non-designers, who were born and educated in Korea and now live in Korea. Participants in the British group were non-designers who were born and educated in UK and now live in UK. In both groups the cultural origin criterion also applied to parents and grandparents, the age ranged from 20 years to 65 years, means 36.8 years and standard deviation 8.45 years. The groups were matched according to age and gender. The participants from the previous experiment differed from the participants in this current experiment.

#### 3.1.3 Materials

As shown in Fig. 1, web pages were designed to show 10 different colors which were presented in sequence. These colors consisted of 5 the most preferred colors by Koreans and another 5 the most preferred colors by English [2]. For instance, the color square in the original example of Fig. 1 was blue as a preferred color of the Korean participants. Each page showed just one color in front of a small gray (50% of Black) background and a phrase or explanation. The latter indicated the meaning of that color and was presented next to that color. The most descriptive phrases were selected from the result of the previous experiment described in Section 2 by a panel consisting of 4 English and 4 Korean members. As result, each color was combined with 4 Korean and 4 English meanings. The web pages are available at http://www.ygassociates.com/gsoonchoi/colours3.

The 10 selected colors were Red (FF0505), Dark Blue (163876), Dark Green (006666), Blue (6484BD), Dark Brown (5D0105), Bright Blue (00B2CC), Pink (F4358D), Bright Green (00ED2A), White (FFFFFF), Yellow (F9EA02). The color-meaning combinations were randomly permutated, so that colors were combined with appropriate as well as inappropriate meanings. For the color blindness test Ishihara's standardized figures were used [7].



Colour Survey C003

<Colour 1> Do you agree with the following description of the colour?

Coventry City FC

**Fig. 1.** Experimental website

### 3.1.4 Apparatus

The web pages were displayed on a Sony VAIO VGN-S28LP Laptop Computer, which was set up in a black box. The box was sufficiently large to include the computer and the head of the participants. In this way participants' perceptions were excluded from any visual effects outside of the computer screen. Measurements of the black box were 65cm width * 40cm height * 85cm depth. The same set-up was used for all participants in Korea and in the UK.

### 3.1.5 Procedure

Each participant was tested for color blindness before participating in the experiment to ensure that they have normal color vision. Participants were required to touch the back of the box with the back of their head. This resulted in a fixed distance between computer screen and head of the participant. The computer screen was set up in an angle of 90 degree to the base and raised to the participant's eye level. The computer was covered with the black box described above. Then the designed web pages were presented in Korean to Korean participants and in English to English participants. After presenting each color-meaning combination, participants were asked to score their appropriateness of the combination on a Likert scale from 1 to 7, where 1 indicates a very inappropriate and 7 indicates a highly appropriate combination.

### 3.2 Results

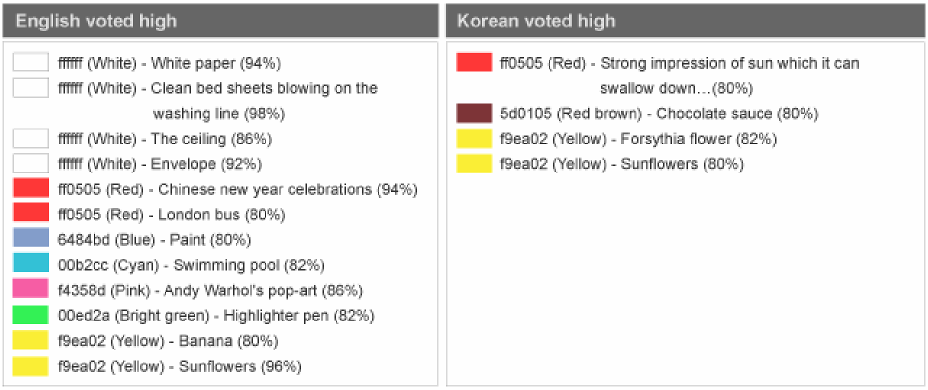
The data have been analyzed using a 2*(10*2*4) mixed ANOVA [8][9] with cultural origin (Korean and English) as unrelated factor and color, cultural origin of the meaning and the actual meaning as related factors. The main effect of color was significant: F = 37.9, p < 0.001. The color by culture interaction was significant: F = 6.5, p < 0.001. The main effect of cultural origin of meaning was significant: F = 31.2, p < 0.001. The culture origin of meaning by culture interaction was significant: F = 55.4, p < 0.001. The main effect of meaning was significant: F = 10.6, p < 0.001. The meaning by culture interaction was significant: F = 4.0, p< 0.001. The color by cultural origin interaction was significant: F = 16.3, p < 0.001. The color by meaning interaction was significant: F = 14.3, p < 0.001. However, the main effect of culture was not significant: F = 3.19, p < 0.077.

In addition to these quantitative findings, there were also a number of qualitative results.

The previous study summarized in Section 2, has been confirmed. English participants responded to meanings that involved artificial objects with high ratings, whereas Korean participants responded to meanings that involved nature-related items with high ratings. For instance, Korean participants provided high ratings for white paper (72%), cloud in the sky (70%) for the color ‘White’ (ffffff) and forsythia flower (82%), warm, soft and downy chicks (74%) and banana (78%) for the color ‘Yellow (f9ea02). In contrast, the most highly voted color-meaning combination by English participants were the following: clown (78%), Chinese new year celebrations (94%), telephone box (70%) and London bus (80%) for the color ‘Red’ (ff0505), Night sky (80%) for the color ‘Navy blue’ (163876) and swimming pool (82%) for the color ‘Cyan’ (00b2cc).

In the current study, English participants responded with a wide range of different ratings. In contrast, Korean participants used a narrower subset of the ratings.

A similar situation was found with those color/meaning combinations that received high ratings. English recognized with a high rating to a wide range of different color/meaning combinations, and Koreans recognized with a high rating to a narrower range of combinations (Figure 2).



| English voted high | Korean voted high |
|---|---|
| ⬜ ffffff (White) - White paper (94%) | 🟥 ff0505 (Red) - Strong impression of sun which it can swallow down…(80%) |
| ⬜ ffffff (White) - Clean bed sheets blowing on the washing line (98%) | 🟫 5d0105 (Red brown) - Chocolate sauce (80%) |
| ⬜ ffffff (White) - The ceiling (86%) | 🟨 f9ea02 (Yellow) - Forsythia flower (82%) |
| ⬜ ffffff (White) - Envelope (92%) | 🟨 f9ea02 (Yellow) - Sunflowers (80%) |
| 🟥 ff0505 (Red) - Chinese new year celebrations (94%) | |
| 🟥 ff0505 (Red) - London bus (80%) | |
| 🟦 6484bd (Blue) - Paint (80%) | |
| 🟦 00b2cc (Cyan) - Swimming pool (82%) | |
| 🟪 f4358d (Pink) - Andy Warhol's pop-art (86%) | |
| 🟩 00ed2a (Bright green) - Highlighter pen (82%) | |
| 🟨 f9ea02 (Yellow) - Banana (80%) | |
| 🟨 f9ea02 (Yellow) - Sunflowers (96%) | |

**Fig. 2.** Voting rate by English/Korean

Finally, English participants showed a higher agreement on high ratings than Korean participants. The agreement on these ratings was in the 80-98% range for English participants and in the 80-82% range for Korean participants.

## 4 Discussion

It was predicted that a meaning generated by Korean participants will attract a high rating from Korean participants and a low rating from English participants. Equally a meaning generated by English participants, will attract a high rating by English participants and a low rating by Korean participants. However, the results showed a discrepancy between the previous and the current study. Although, in the current study all related factors had significant effects, the unrelated factor of culture did not. In contrast, the previous study showed a clear preference of the English participants for meanings related to artificial objects, whereas the Korean participants preferred nature-related objects.

The reason for the discrepancy between the two studies can be seen in the different design. The previous study used a free association task. The participants had for each presented color to state what first came to their mind. In contrast, the current study utilized a recognition task. Participants were presented with a color and a meaning at the same time. So their judgment was based on the extent to which the color/meaning combination makes sense. Whereas the free association task depended on the cultural context of the participant, the recognition task was based on the participant's cross-cultural knowledge. In this situation, a cultural meshing may take place, where meanings of a foreign culture become more dominant. For instance, English participants rated the color red and the Chinese New Year celebrations high, but rated the color 'Cyan' and Jeju Island low, because they had the cross-cultural knowledge available in the first case but not in the second. This result is relevant for cross-cultural chance discovery in general. It suggests that in creative mental tasks people draw from their own cultural background. However in evaluation tasks, where they attempt to make sense of a given scenario, they consider cross-cultural knowledge as well. In this sense, the free association task is more sensitive for identifying culture-based new chances than the recognition task. This may be particularly relevant if members of the English language group are involved, as many countries have experienced a strong culturally dominant influence of the American culture. For instance, in a cross-cultural group that generates scenarios for chance discovery [10], it can be expected that group members generate their own scenarios within their own cultural context, but evaluate the scenarios of other group members by drawing from their cross-cultural knowledge. They may then accept scenarios, which are culturally less sensitive.

The results also confirm the previous result, which indicated that English participants responded with high ratings to meanings based on artificial objects and Korean participants responded with high ratings to nature related meanings. This difference can be explained within the historically situated knowledge. England moved much earlier from an agricultural society to an industrial society than Korea. Therefore English participants may led to a larger extent have lost their contact to nature.

Also the differences in the range of color/meaning combinations that were rated highly can be explained with differences in the historically situated knowledge. During the imperialistic period of British history, Britain did not just occupy other countries but also traded with them and brought their cultural artifacts home. This development led to openness to other ideas and cultural concepts. Later many inhabitants of these countries won the right to live in the UK. This again led to openness to and a partial understanding of foreign influences. In contrast, Korea until the 20th century was a rather closed society. Therefore the view of what cultural meanings make sense may be more restricted. This restriction may also have led to a lower agreement in high ratings.

Differences in the societal make-up could also explain why Korean participants preferred narrower ratings than English participants who used the entire scale available to them. Korea is a mostly collectivist society, which may have led to a larger agreement among participants, whereas the individualism in the English society may lead to more diverse views.

It should however be noted that the current study presented colors and meanings in isolation. However, even in a single Web page colors are presented in many different contexts. Therefore the next study in the current research program will investigate the constraints that a given context imposes on the interpretation of the meanings of colors.

**Acknowledgments.** The first author gratefully acknowledges the support of the Arts and Humanities Research Council in the UK.

## References

1. Hu, J., Shima, K., Oehlmann, R., Zhao, J., Takemura, Y., Matsumoto, K.: An Empirical Study of Audience Impressions of B2C Web Pages in Japan, China and the UK. Electronic Commerce Research and Applications 3(2), 176–189 (2006)
2. Choi, G.S., Oehlmann, R., Dalke, H., Cottington, D.: Cross-cultural comparison of color preferences between English and Korean subjects. In: Proceeding of Social Intelligence Design, Trento, Italy (2007)
3. Berry, J.W., Poortinga, Y.H., Segall, M.H., Dasen, P.R.: Cross-cultural psychology. Cambridge University Press, Cambridge (2002)
4. Brislin, W.R., Lonner, J.W., Thorndike, M.R.: Cross-cultural research methods. Wiley, New York (1973)
5. Hallway, W.: Methods and knowledge in social psychology. In: Hallway, W., Lucey, H., Phoenix, A. (eds.) Social psychology Matters. The Open University, Milton Keynes (2007)
6. Choi, G.S., Oehlmann, R., Dalke, H., Cottington, D.: Discovering color semantics as a chance for developing cross-cultural design frameworks, Chance Discovery, Salerno, Italy (2007)
7. Ishihara, S.: Ishihara's Tests for Color Deficiency. Kanehara Trading Inc., Tokyo (2001)
8. Gray, C., Kinnear, P.: SPSS for Windows made simple. Psychology Press, Hove (1994)
9. Brace, N., Kemp, R., Snelgar, R.: SPSS for Psychologists, 3rd edn. Palgrave-Macmillan, Basingstoke (2006)
10. Oehlmann, R.: The Function of Harmony and Trust in Collaborative Chance Discovery. New Mathematics and Natural Computation 2(1), 69–84 (2006)

# A Cross-Cultural Study on Attitudes toward Risk, Safety and Security

Yumiko Nara

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

**Abstract.** In this paper, the author aims to examine the status quo of people's attitude toward risk and safety as well as the differences between risk attitudes in Japan, USA and China. The social survey was carried out in February and March of 2008. It used questionnaires to obtain data of people: male and female, 20-69 years, in each country, using random sampling. The survey has clarified the status quo and differences about perception toward 19 items of risk, effects of safety perception and risk image based on people's anxiety. It has been found that Japanese people fear most strongly the bad influence of life risks. The effect of risk image on anxiety was relatively strong in China and Japan, the effect of safety perception was stronger in the US.

**Keywords:** risk, safety, security, risk image, comparative study, questionnaire, Japan, the United States, China.

## 1 Introduction

It has been argued that today is the era of risks. Actually the environment that surrounds us is full of various risks.  The concept and definition of risk vary depending on the field of study and the researcher. Risks are understood as 'possibilities,' 'uncertainties,' 'deviations of expectations and results,' 'probabilities,' and 'expected losses,' which are contained in hazards, as well as obstacles and other undesirable phenomena which jeopardize the safety of life, health, assets, socioeconomic activities, and the natural environment. Despite such variations, the true nature of risks is invariably understood as the 'occurrence of undesirable phenomena' that may inflict 'harmful influences.' Thus, everyday life risks can be classified as material risks, human risks, and indemnity-liable risks, depending on the type of exposure. They can also be classified by the conditions that give rise to risks, such as risks related to highly advanced science and technology, environmental problems, consumers' lives and products, health and medical problems, and disasters.

In order to examine actual conditions, backgrounds and solutions of risks, various studies have been conducted in the fields of social sciences and natural sciences. The important tendencies observed in recent risk studies can be characterized in terms of the following views; 1) They are focusing on cross cultural studies on risk under the tendency of globalization, 2) They are focusing on the viewpoints of citizens (living

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 734–741, 2008.

subjects, consumers, ordinary people; non-specialists/scientists of risk) based on the necessity of risk governance/communication.

As for the former view, there are significant studies of risk management and policy, disaster prevention, public health, economics and so on [Okada 2004; Kan & Chen 2003]. From the view point of the latter, the importance of citizen's standpoints have been emphasized in theoretical approaches and in the preliminary or secondary data [Kikkawa 1999].

There are some empirical studies based on both viewpoints [Hirose & Slovic 1994; Kleinhesselink & Rosa 1991, Tsuchida & Pergar 2006], however, most of them treat certain kinds of risk in certain populations. This study is trying to approach everyday life risks with these two points of view, i.e., the author aims to examine the status quo of people's attitude and coping toward various kinds of everyday life risks as well as the differences between Japan, USA and China using questionnaire survey data obtained by random sampling in all three nations.

This study also aims to realize structural research. Actually there are worthy studies which focus on risk perception itself [Slovic, 1987]. Others that are also important examine each relationship between risk perception (or risk coping) and income, age, sex, occupation, view of fate or nature and so on [Nakamura & Sekiya 2004; Flynn, Slovic & Merts 1994, Lazo, Kinnell & Fisher 2000]. Based on these worthy studies, the author is trying to grasp whole structural conditions and relationships among attitude toward risk, coping with risk, various living attributes, and total evaluation toward living.

## 2 Analytical Framework of Study

The analytical framework of this study is shown in Figure 1. The whole framework of this study consists of three parts; 1) attitude toward risk, 2) coping with risk and 3) evaluation of living. The main hypotheses of this study are as follows; Hyp.1 'People's total evaluation of their living depends on the extent of adaptation between their attitude toward risk and coping with risk' and Hyp.2 'the actual condition of risk attitude/coping and relationships among various variables are different in Japan, USA, and China'.

The part of 1) attitude toward risk has yet two more components – risk perception and risk philosophy. Here are two sub-hypotheses; SubHyp.1 'People's risk perception is affected by risk philosophy' and SubHyp.2 'Their perception of safety/anxiety is effected by the perception of risk and risk image'. As for the part of risk coping, it also has three components; 1) resource, 2) risk management process, and 3) risk management philosophy. About risk coping, there are two more sub-hypotheses; SubHyp.3 'People's condition of the risk management process is influenced by the quality and quantity of their resource', and SubHyp.4 'Their condition of the risk management process is affected by their risk management philosophy'.

Among the above analytical framework, this paper is focusing on risk perception, i.e. to grasp the actual condition of safety perception, risk image, and security/anxiety perception. Moreover Hyp.2 and SubHyp.2 will be examined.

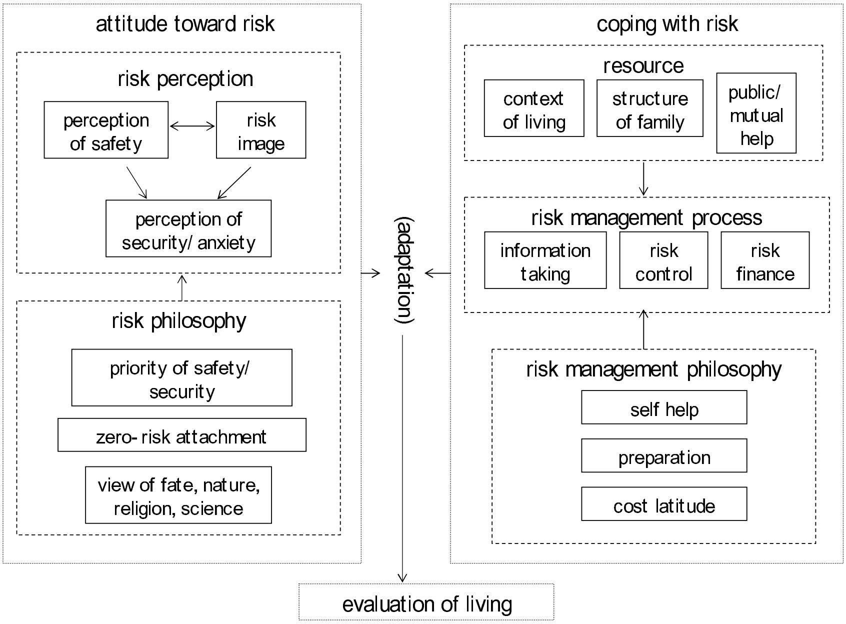


**Fig. 1.** Analytical framework of the research

# 3 Method

## 3.1 Indexing of Variables: Focusing on 'Safety', 'Security' and Risk Image

The framework shown in Fig.1 aims to clarify the people's condition around safety and security. Safety is defined by the situation in which the risk level if objectively evaluated is sufficiently low to be accepted. Security is grasped as the situation that people subjectively think risk is low.

The specialists/scientists produce safety items when they research and manage risks. On the other hand, ordinary people/citizens cannot be free from perception of security. Social psychologists have pointed out that risk perception between citizen and specialists/scientists is different [Okamoto 1995]. Risk is evaluated with its frequency and severity. However people/citizens –not the specialists/scientists – often overestimate risk with some reasons; adding risk image, using various and compound endpoints, using short information as well as literacy, rejecting unacceptable risk, and being affected by his/her living attributes [Nara 2007].

In the case of members of the general public, risk perception is synonymous with understanding uncertain phenomena. This understanding is based on an integrated set of criteria, including frequency and intensity, anxiety and fear, optimism, benefits, and acceptability, from subjective viewpoints. Therefore, risks are perceived differently depending on each individual's limitation of information-processing ability, surrounding situations, and attributes, such as past risk career, age, sex, occupation, and academic background.

On the other hand, in the case of risk assessment and management organizations, including experts, companies, and administrative organizations, risks are analytically perceived based on estimations of probability, quantitative measurements of losses, and calculations of costs and benefits. Accordingly, there is a great discrepancy in risk perception between individuals and organizations. Risks perceived by the former are subjective risks, while those perceived by the latter are objective risks.

Especially image of risk in people would make significant effects on her/his estimation. Based on these perspectives, the author tried to grasp the status quo of people's perceptions toward safety, risk image, and security/anxiety. Furthermore the relationships among these three concepts should be examined.

Indexes of main variables are shown below. Among these indexes, risk image is based on the pre-standardized scales, the others were made by the author specifically for this study.

### Items of risk

As for the substantial indication of variables, the following nineteen risks which would be happen in everyday life were presented; a) Earthquake, b) Traffic accident, c) Fire, d) Cancer, e) Food laced with foreign substances or chemical substances, f) Involvement in a crime (as a victim), g) Illness and injury, h) Decrease in income, i) Decrease in assets, j) Financial difficulties after retirement, k) Global warming, l) Health hazards from genetically modified food, m) Side effects of drugs, n) Accidents at nuclear power plants, o) Internet scams, p) Leaks of personal information on the internet, q) Defamation of people on the internet, r) Involvement in sexual crimes on the internet, s) Computer viruses.

### Perception of security/anxiety

Concerning about each risk above the nineteen items, the level of security/anxiety perception is obtained with the following question;

1) How anxious do you feel regarding risks listed below in your daily life? Use a number from 1 to 6, where 6 means "very anxious" and 1 means "not anxious at all."

### Perception of safety

The level of safety perception is obtained based on two elements; 2) the probability of a risk and 3) the severity of the consequence of a risk, with the following questions;

2) How great is the chance that these things happen to you? Use a number from 1 to 6, where 6 means "definitely to happen" and 1 means "will never happen." (Circle one for each risk)

3) Suppose the following things happened to you. How much do you think you would be damaged? Use a number from 1 to 6, where 6 means "Seriously damaged" and 1 means "Not damaged at all." (Circle one for each risk)

### Risk image

Risk image is an image (whole and integrated impression) which an individual holds toward risk. Risk image is formulated with the peculiarity and content of that risk. The level of risk image is obtained based on the scale of the risk image components described by Slovic [Slovic 1987]. In this study it is composed of one factor of the unknown risk; 4) the clarification level of each risk by the specialists and 5) the knowledge level of a person (participant of the study) toward each risk, and another factor of the dread risk; 6) controllability over risk. Substantial questions are the following:

4) To what extent, do you think the following risks are scientifically made clear by professionals? Use a number from 1 to 6, where 6 means "very clear" and 1 means "not clear at all."

5) How much knowledge do you think YOU have concerning these risks? Use a number from 1 to 6, where 6 means "a lot of knowledge" and 1 means "no knowledge at all."

6) To what extent do you think you can avoid these risks on your own before they happen? Use a number from 1 to 6, where 6 means "Can avoid completely" and 1 means "Can't avoid at all." (Circle one for each risk)

### 3.2 Outline of the Survey

This survey was carried out with the following frame; 1) Population and subjects of survey: [J] & [US] & [C] male and female, 20-69 years old, all parts of the country. 2) Measure for Questionnaire: [J] & [US] send and return it by mail, [C] by telephone (CATI: Computer Assisted Telephone Interview). 3) Sampling ledger: [J] NOS list, [US] GfK list, [C] RDD (telephone number ledger). In three countries, subjects were sampled with the simple random sampling method according to a sex and age according to population percentage. 4) Number of the Useable Samples: [J]1,050s, [US]509s, [C]1,000s. 5) Survey period: [J] 2008 Feb13-Feb29, [US] 2008 Feb23-Mar28, [C] 2008 Feb18-Mar6. 6) Investigation implementation organization: [J] Nippon Research Center, [US] GfK Custom Research North America, [C] Chinese Academy of Social Sciences.

Basic attributes of respondents are as follows; Gender: female 54.6% and male 45.4% in Japan, 50.9% and 49.1% in US, 50.0% and 50.0% in China. Age: 20-29 years old: 14.1 %, 30-39: 23.5 %, 40-49: 21.1 %, 50-59: 20.3 %, 60-69: 20.9 % in Japan (average 45.65 years old), 20-29 years old: 22.6 %, 30-39: 22.2 %, 40-49: 22.0 %, 50-59: 18.6 %, 60-69: 14.7 % in US (average 42.65 years old), 20-29 years old: 24.0 %, 30-39: 30.0 %, 40-49: 23.0 %, 50-59: 16.0 %, 60-69: 7.0 % in China (average 39.55 years old) .

## 4 Results and Discussion

### 4.1 Status Quo of Attitude toward Risk

About risk perception, Japanese people fear most strongly the bad influence of the life risks of the 19 items, although the presentation of the descriptive distribution of the reply results is omitted due to a lack of space. Another result shows that Japanese people's self-perception on scientific elucidation and individual knowledge on risks is relatively low. It has been stated by other researchers (e.g. Hofstede 1991; Mizushima 2002) that Japan is the society in which people feel a high dread, and in this study the same result was obtained.

For the perception of security a MANOVA (multiple analysis of variance) was conducted. It showed the following significant differences by country: 1) anxiety related to risk – the Japanese feel great anxiety toward all 19 risks (Wilks $\lambda^2$= .421

p<.001). There are the following tendencies of the result; Japan > China > US about the risk items except related to the Internet, and Japan > US > China about the risks of the Internet.

On the perception of safety, the following results are obtained; 2) the probability of risk -- Japan > US > China (Wilks $\lambda^2$= . 334 p<.001), and 3) the severity of risk -- Japan > US > China (Wilks $\lambda^2$= . 547 p<.001). With regard to substantial risk, Japanese people estimated a higher probability and severity toward all 19 items of risks. The possibility of occurrence of big earthquake, for example, is definitely quite highest in Japan. However the Japanese find high probability even toward other risks which occur with almost the same or less probability in Japan compared with the other two countries, like cancer, traffic accident, crimes on the Internet and so on.

As for the risk image, the result of 4) the clarification level of each risk by the specialists is significant (Wilks $\lambda^2$= . 721 p<.001) with the following tendencies; China > US > Japan about the risk items except related to the Internet, and US > China > Japan about the risks of the Internet. Moreover the result of 5) the knowledge level of each risk of a person shows that Japanese do not think they get sufficient knowledge about risks (Wilks $\lambda^2$= .701 p<.001) with the following tendency: US > China > Japan. Lastly the result of the controllability of risks, the tendency of China > US > Japan (except Internet risks) and US > China > Japan (Internet risks) were observed (Wilks $\lambda^2$= .655 p<.001). These results indicate that Japanese people feel the unknown of risks more strongly than the Americans and Chinese do. Moreover, as on the dread of risks, the results also indicate that Japanese people feel the dread of a risk strongly.

### 4.2 Relationships between Perception of Security, Safety and Risk Image

In order to examine the relationships between people's perception of security/anxiety, safety and image of risk, a multiple regression analysis was conducted. The dependent variable here is the perception of security/anxiety, independent variables are the perception of safety (the probability of risk and the severity of risk consequence) and risk image (the clarification level of each risk by the specialists, the knowledge level of each risk of a person, and controllability over risk).

The result is shown in Table 1. In this table the scores of β (standardized regression coefficient) for anxiety of each risk (here are five items, however, the others show almost the same tendencies) are observed. It indicates that these independent variables generally affect the level of people's anxiety. However, the size of the effect varies depending on the independent variable as well as the country.

The score of β of safety perception (frequency and severity of risk) is large in Japan and the US. This tendency is remarkable in the United States. The influence of the risk image is observed in Japan; on the other hand, US's influence of the risk image is small or even insignificant. In China, the influence of the knowledge level of each risk of a person is greater than the influence of safety perception (frequency and severity) unlike Japan and the United States. After adding the main basic attributes (age, annual income, and sex) as independent variables, almost the same tendency is obtained. However, the effect of the annual income is considerably large in China

**Table 1.** Result of multiple regression: β (standardized regression coefficient) for anxiety of each risk

| | Earthquake | Cancer | Decrease in income | Global warming | Leaks of personal information on the internet |
|---|---|---|---|---|---|
| Japan | | | | | |
| probability | .314 *** | .286 *** | .239 *** | .203 *** | .206 *** |
| severity | .243 *** | .202 *** | .374 *** | .371 *** | .327 *** |
| Clarification by specialists | .021 ns | −.020 ns | −.015 ns | −.041 ns | −.056 * |
| Knowledge of a person | | .108 *** | .066 * | .192 *** | .138 *** |
| controllability | .009 ns | −.047 ns | −.053 ns | .032 ns | −.099 ** |
| | (R2= .212) | (R2= .171) | (R2= .299) | (R2= .327) | (R2= .280) |
| USA | | | | | |
| probability | .442 *** | .410 *** | .288 *** | .326 *** | .265 *** |
| severity | .091 * | .208 *** | .315 *** | .325 *** | .341 *** |
| Clarification by specialists | −.012 ns | .006 ns | .053 ns | .156 *** | .087 * |
| Knowledge of a person | .076 ns | .016 ns | .118 ** | .052 ns | −.056 ns |
| controllability | .004 ns | .085 * | −.077 ns | .025 ns | .026 ns |
| | (R2= .247) | (R2= .258) | (R2= .311) | (R2= .451) | (R2=.256) |
| China | | | | | |
| probability | .079 * | .054 ns | .128 *** | −.008 ns | .197 *** |
| severity | .138*** | .149 *** | .155 *** | .229*** | .286 *** |
| Clarification by specialists | .030 ns | .253 *** | .181 *** | .039 ns | .078 ** |
| Knowledge of a person | .171*** | .158 *** | .166 *** | .183 *** | .183 *** |
| controllability | .055 ns | −.097 * | .018 ns | .012 ns | .004 ns |
| | (R2=.072) | (R2= .153) | (R2= .152) | (R2= .107) | (R2= .324) |

* p< .05 ** p< .01 *** p< .001

compared with other countries. In Table 1, the scores of R $^{2}$ of China tend to be smaller than those of Japan and the US. Therefore the effects of other independent variables have to be examined to eliminate the residual.

## 5 Conclusion

This study has tried to clarify the status quo of risk perception and relationships between perception of safety/anxiety, perception of safety and risk image. The results can be summarized as follows: Hyp.2 is verified at least with respect to the perception of risk, i.e. the actual condition of risk perception and relationships among various variables are different in three courtiers. SubHyp.2 is verified. People's perception of safety/anxiety is affected by the risk perception and risk image. In this paper, only a part of the entire analytical model has been treated. It is expected that various factors form anxiety. For instance, it seems that the view on fate, religion, nature and science have exerted influence on risk perception. Moreover, people's feeling to be relieved under the perfect condition with zero-risk should be examined. Further future works is to clarify the relationship between attitude toward risk and coping with risk as well as the effect of these on the satisfaction with the entire living situation.

## Acknowledgement

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

## References

Flynn, C.K., Slovic, P., Merts, C.K.: Gender, race & perception of environmental health risks. Risk Analysis 14, 1101–1108 (1994)

Hirose, H., Slovic, P., Ishizuka, T.: A comparative research of risk perception on US-Japan college students. Research in social psychology 9(2), 114–122 (1994)

Hofstede, G.: Cultures and Organizations: Software of the Mind. McGraw-Hill, New York (1991)

Kan, H., Chen, B.: A Case-crossover Analysis of Air Pollution and Daily Mortality in Shanghai. Journal of occupational health 45(2), 119–124 (2003)

Kikkawa, T.: Risk Communication. Fukumura-shuppan (1999)

Kleinhesselink, R., Rosa, E.A.: Cognitive Representation of Risk Perception: A Comparison of Japan and the United States. Journal of Cross-cultural Psychology 22, 11–28 (1991)

Lazo, J.K., Kinnell, J.C., Fisher, A.: Expert and layperson perception of ecosystem risk. Risk Analysis 20, 179–193 (2000)

Mizushima, K.: Risk and Japanese Society. Japanese Journal of Risk Analysis, 1–14 (2002)

Nakamura, I., Sekiya, N.: Safety Perception of Japanese. Basic Research Project on Nuclear Power and Safety (2004)

Nara, Y.: Security, Safety and Risk Management. Risk and Insurance Management 38, 115–128 (2007)

Okada, N.: Urban Diagnosis and Integrated Disaster Risk Management. Journal of Natural Disaster Science 26(2), 49–54 (2004)

Okamoto, K.: Introduction to Risk Psychology: Human Errors and Risk Images. Science (1995)

Tsuchida, S., Pergar, K.: Female perception of risk with regard to cultural background. Bulletin of the Faculty of Sociology 37(3), 39–53 (2006)

Slovic, P.: Perception of Risk. Science 236, 280–285 (1987)

# Insight or Trial and Error: Ambiguous Items as Clue for Discovering New Concepts in Constrained Environments

Jun Nakamura[1,2] and Yukio Ohsawa[1]

[1] The University of Tokyo, Bunkyo-ku, Tokyo, 113-0033, Japan
[2] Bearing Point, Chiyoda-ku, 100-6223, Japan

**Abstract.** As our life becomes more sophisticated, our ability to create concepts is indispensable. Based on our concept formation model, which focuses on ambiguous items that cause a change of one's decisions, we developed a web-based creativity support system. This is intended to guide users in the categorization of words in analogical reasoning by forming batches of words to discover new concepts. 20 junior high school students participated in an experiment, where concept formation is expected to be induced by an ambiguous interpretation of presented information under constraints. The result provided some evidence that an existence of ambiguous items corresponds to a diversification of a thought process either as *insight* or *trial & error* in categorizing words. The result also showed that the reconstruction of groups for including ambiguous items into some groups accelerates the creation of new concepts for better interpretation.

**Keywords:** Creativity, concept, ambiguity and constraints.

## 1 Introduction

Consider a business manager in some liquor production company with decreasing revenues, who may investigate various ways of improving sales, such as discounting products, increasing the sales force etc. In this situation during a visit of your factory, the manager may note employees who fit the cap to the liquor bottle in the production line. He may become aware that the cap is not liquor itself but a supplementary item, and simple part of the bottle.

One of the scenarios is that this cap could be interpreted as an extreme happy time for consumers when they begin to open the bottle at night. Given such a scenario, the manager may become aware of the necessity to provide a new concept of service by offering a pleasant time and place, where the liquor can be enjoyed in a happy and relaxed atmosphere, where it is sold at a discounted price.

This is an example for the necessity of using analogies to become aware of new concepts. This thought process is important for identifying the characteristics of new concepts, when the concepts cannot be seen physically such as displayable products.

We classified two notable types of study related to this concern, one focuses on analogical reasoning to support creativity, and the other on analytical references to evaluate human behavior.

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 742–749, 2008.

Evidence from cognitive science and psychology suggest that people can ascribe a variety of contextual images to one given word. Sometimes the meaning is changed with a interpretation. This allows discovering a new interpretation [1]. This ambiguity of word meanings is reflected in a perceptual shift that can be characterized either as a subjective or objective transfer [2,3]. Here we focus on analogical reasoning processes, which are essential for the creative formation of new concepts [4,5]. Analogical reasoning has been modeled in the structure mapping theory [1,6], which allows the transfer of knowledge between different domains. This cross domain transfer from the base (given words) to the target (concepts) enables the generation of new creative designs and has been considered to be a discovering process. Generally the formation of an analogy requires a hint about the relationships between base and target problem. [7]. Without such a hint most people are not able to form the analogy.

However, these studies have not considered words that *cannot* or *are difficult* to be interpreted for new categorization. We, therefore, focus on such words as "*ambiguous items*", and this is the main issue of this paper.

Equally relevant are empirical methods to analyze human cognitive behavior. Creative activity can effectively be supported by discovering combinations of concepts, where the combinations are comprised of keywords [8].

This method has been described on an algorithmic level and the approach taken in this paper benefited from it. However this paper will introduce an experimental tool to observe human cognitive behavior. As a tool, the newly developed environment is based on the metaphor of word cards. There are several methodological approaches using cards, e.g. the CRC approach and Q-methodology. The CRC approach [9,10], which facilitates the process of discovering real world objects, is designed for object oriented modeling according to predefined scenarios, whilst our target is to create the actual scenarios by combining word cards. The Q-methodology [11] is a way to reveal subjective structures, attitudes and perspectives by using cards. The objective is similar with our paper, but it is based on factor analysis involving the rank-ordering of a set of statements ranging from *agree* to *disagree*. The intention therefore is not to explore the thought processes of creativity.

We assume that reviewing an individual's own interpretation, if affected by the presence of ambiguous items, might lead to the discovery of new concepts.

In the following section, the concept creation model will be presented. This model has been used in the subsequent experiment (Section 4). The results of the experiment in the light of the model will be discussed in Section 5.

## 2 Concept Creation Model

The concept creation model, as shown in Figure 1, is based on the idea that outer knowledge contributes to creativity [12]. As the constrained environment motivates an individual's creativity [13], a rule was designed stating that the subject must not leave any of the given word cards without being categorized. That is, all the given word cards must finally be categorized in some clusters.

We assume that people tend to address any items that are easy to categorize first and leave out some items, which are difficult to categorize. Under the constrained rule, we expect that subjects will be sufficiently activated to discover new concepts by recombining clusters into new groups, as indicated in Figure 1.

The state of uncertainty corresponds to an impasse [14], which is caused by interference from preconceived ideas. But the constrained rule and the instruction to the game might enable subject to overcome the impasse.

We assume that if one does not leave any items out (i.e., uncategorized items do not exist), subjects might obtain new *insight* by looking over all of the given words during the entire process of the experiment. If subjects leave items out (ambiguous items), they might be obliged to cluster these uncategorized items with a *trial & error* strategy.

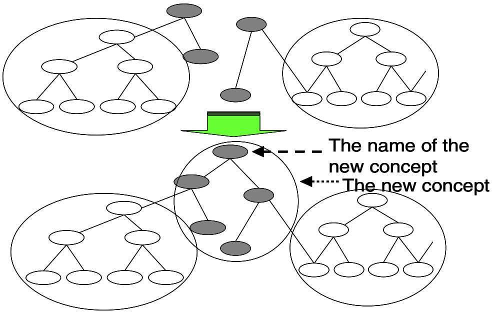


**Fig. 1.** Reconstruction of groups by adding items that were left out. The progress is from two clusters with four uncategorized (uncertain) items in the upper image to three clusters with no uncategorized items in the bottom image.

**Fig. 2.** Students enjoy playing an analogy board at junior high school

# 3 Experiment

## 3.1 Method

### 3.1.1 Design

Players can consider the structure by relocating word cards on the screen to embrace a wider perspective (1). If the player discovers a common concept corresponding to a cluster, he colors the nodes in the cluster and writes the meaning of the words according to the player's interpretation (2). The player then assigns a concept name to the cluster, the word cards of which are painted in the same color (3). If some word cards are not categorized in any cluster (ambiguous items) (4), the player restructures a cluster [15] and finds a new concept that allows for the reconstruction of the clusters (5). The player continues to work through the cards until all word cards are categorized into clusters (6).

All the words is simple nouns, but they include ambiguities [2], as the word *Lincoln* may mean a president of the United States, the name of a place or the name of a vehicle type. We expect that this ambiguity triggers participant's analogical thinking in their mental space [16], and that this will provide a clue for a sufficient link with other words and hence enable the discovery of combined concepts.

### 3.1.2 Participants

Twenty students at junior high school have enjoyed playing the analogy board, as shown in Figure 2.

### 3.1.3 Materials

Based on the concept creation model explained above, we constructed the interface shown in Figure 3. The player is presented with 20 words (items), each in a small square node. Initially, the words are randomly placed, with no objective relevance (Figure 3-A). The most of the words are categorized by filling the meaning on the card, except ambiguous items (Figure 3-B). A group is reconstructed by involving ambiguous items (Figure 3-C).

We prepared a 20-word card set: strawberry, baseball, internet etc. This set can be flexibly changed according to the experimenter's objectives.

The system architecture is rather simple. A network environment with a browser is sufficient for the player to engage with the system. The system provides motion capture logs to the server each time the player performs any action.

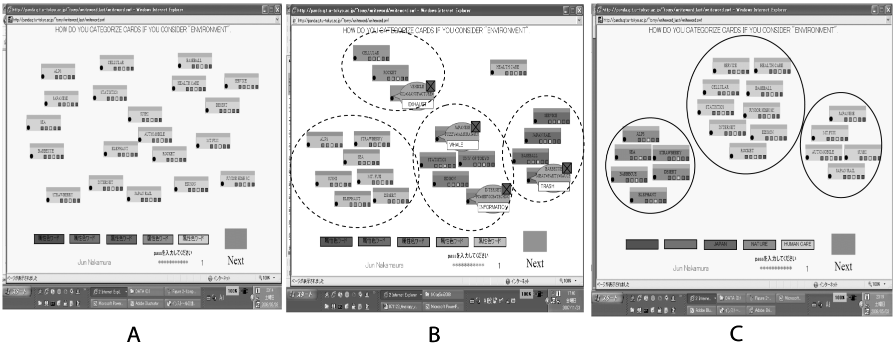


**Fig. 3.** Screen image of the developed environment, where (A) 20 word cards are located randomly at the beginning, (B) 4 clusters with one item left aside. The player clicks on the orphan card to view a balloon, where the player writes the meaning of the word, as interpreted by the player. and then (C) all items are categorized into three clusters, by reconstructing combinations.

### 3.1.4 Procedure

The experiment was conducted twice without break, the first time with no special instruction and the second time with the instruction to focus on a social issue, such as "consider environmental issues" to avoid the impasse [14].

## 3.2 Results

### 3.2.1 The Analysis Method

From the various motion-related data logs generated by the system, we abstracted logs of actions according to the following analysis method.

First, there are 3 types of action (to drag the item, to color the item and to write the meaning). We focus on mainly two actions, one is to color the item, and another is to write the meaning of the item. These two actions can be considered directly related to awareness of the target (i.e., concepts).

Second, we define the ambiguous item to be extracted from 20 word cards.

Third, we visualize actions in a time frame in order to compare the group of players with and without ambiguous items.

Here we define the cognition of clusters and meaning, and ambiguous items.

– **Cognition of clusters and meaning**

Coloring cards is understood as the cognitive action of discovering a cluster. The cognition of being aware of a cluster is defined by Eq. (1), where the experiment time is normalized by 50 as t={1,2, ・・・ ,50}, and the number of actions required to color the card $C_{\mathrm{i}}$ is defined as $g_i$ （t）

$$G = \sum_{i=1}^{20} \sum_{t=1}^{50} g_i(t) \tag{1}$$

The cognition of being aware of a meaning is captured in the same manner as the cognition of a cluster. Eq. (2) defines the cognition of being aware of a meaning, where the number of actions required to give meaning to an interpretation is expressed as $h_i$ （t）.

$$H = \sum_{i=1}^{20} \sum_{t=1}^{50} h_i(t) \tag{2}$$

– **Definition of ambiguous items**

If players don't have any impasse, there are no items left. If players left some items, we refer to them as ambiguous items, which shall be extracted in the following manner.

We defined here a word card set as $\{C_1, C_2, \cdot\cdot\cdot\ C_{20}\}$. The timing when G' attain 10% in Eq (3), is shown as $t^d$. In this regard, we can note $C_{\mathrm{i}}$ (possibly plural number) as ambiguous items, where $g_i$ (t) > 0.

$$G' = \sum_{i=1}^{20} \sum_{t=d}^{50} g_i(t) \Big/ G \tag{3}$$

The time frame will now be considered for more accurate identification of ambiguous items, because a certain amount of time is at least needed to manage the ambiguities. The above calculation intends to identify the word cards, which are used within the last 10% of all actions. That is, these word cards are left until the last stage of an experiment, or are colored repeatedly.

The timing of the largest number of coloring actions during a complete process shows that players develop their entire strategy by forming concepts, because this action involves the recognition of concepts based on sets of colored word cards. We define here the timing of this mode as $t^m$. We have to pay attention to the length of the time gap between the mode ($t^m$) and the final coloring action ($t^d$), where $t^d. - t^m$ >0. The length can be logically a maximum of 49 ($t^m$ =1, $t^d$ =50) and a minimum of 1

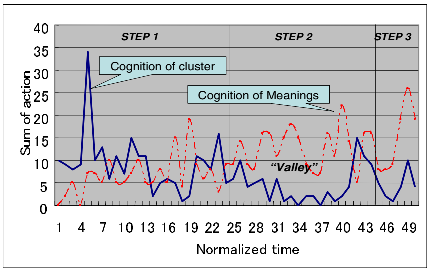
($t^m$ =49, $t^d$ =50). Depending on the length, there are two types of ambiguity, (a) the longer its length is, the more ambiguities are realistic, and (b) the shorter its length is, the more ambiguities are unrealistic. This means that if one leaves some card for a long time the player is struggling in a *try & error* approach to group the uncategorized items and this situation seems to be ambiguous. But if one leave some cards for a very short time it is possible that players temporally place the card aside to consider the grouping of other cards. They appear to consider concurrently both the card that is left aside and other actions with their perceptive *insight* to look ahead over all word cards. That is the reason why players leave the card for a shorter period. The word card left aside, therefore, might not to be ambiguous.

### 3.2.2 Summary of the Result Data and Graphs

We identified ambiguous items for the first game in accordance with the analysis method in Section 3.2.1.

As a result of the procedure, there are 11 players with ambiguous items and 8 players without ambiguous items, for the case that the length between $t^d$ and $t^m$ is 10, to distinguish whether ambiguous items exist or not.

In order to compare the cognition of clusters and meanings for both groups, actions are shown in Figure 4 and 5, in accordance with Eqs.(1) and (2).



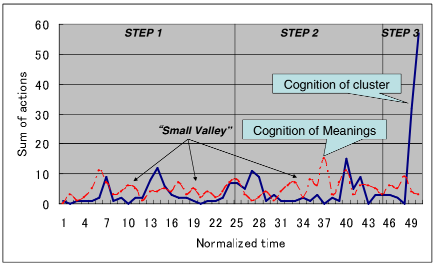


**Fig. 4.** Action related to the cognition of clusters and meanings (for the players with ambiguous items), where the length between $t^d$ and $t^m$ is more than 10

**Fig. 5.** Action related to the cognition of clusters and meanings (for the players without ambiguous items) where the length between $t^d$ and $t^m$ is less than 10

The cognition of clusters, as colored items, is directly related to the awareness of discovering concepts. A t-test has been conducted that showed that the difference between the means of both groups (Table 1) was significant (t=2.110, df=17, p<0.05).

**Table 1.** Result of t-test for the difference of population mean

| | Step1 | Step2 | Step3 | ALL |
|---|---|---|---|---|
| Means | ***2.40751*** | -0.38570 | ***-2.19515*** | 0.07036 |
| Variance | 2.05935 | -0.24960 | -1.89685 | -1.28884 |
| Maximum | 1.86639 | -0.37690 | ***-2.60046*** | -1.40441 |
| Sum of actions | ***2.40751*** | -0.38570 | ***-2.19515*** | 0.07036 |

Let us make our interpretations. Figure 4 shows that coloring (i.e. clustering) of items increased preceding meaning. After this, meaning assignment remained at a high level, whereas clustering activity faded. Thus, in the mid-range of playing time (*step2*), players tended to color some items but left others uncolored. Then they began to write word meanings. During this time, players considered where uncolored cards (ambiguous items) should be integrated or named a new concept rearranging a set of cards that were redefined in meaning, where individuals created new concepts while rearranging ambiguous items. This confirms that ambiguity leads to new concept creation during as you can see "valley" in *step 2* in Figure 4.

On the other hand, in case of players without ambiguous items, we could observed two further distinctive phenomenon, i.e., (a) coloring alternates meaning in most of the process from the beginning, and (b) coloring increases extremely at the final stage. This implies the preparation for final coloring action is taken place from the beginning carefully with "smaller valley".

In either case with the valley that varies in size, the actions in *step 2*, which are observed as the "valley" in Figure 4 and 5, represent one of the core findings of this study. The result of t-test in Table 1 shows that there is no apparent difference between the two parties in *Step 2.*

## 4 Discussion

Whatever ambiguous items are presented, players have to respond to create concepts in the experiment. But as far as the thought process is visualized through the experiment designed according to the concept creation model, ambiguity tends to be induced by words that are difficult to group, i.e., words that are left uncategorized in the experiment, and the reconstruction of groups to include the uncategorized words accelerates a better interpretation by redefining the meaning. In detail, we observed players who considered their individual contexts in the mid-range of the thought process, which is referred to as "valley" in this paper, and concentrate on assigning consistent meaning to words within the cluster, prior to completing the composition. In this regard, the "valley" can be considered an exploratory cognitive process [17, 18].

Depending on the existence of ambiguous items, the thought process for discovering new concepts diversifies especially in the beginning phase and the end phase. It was considered that the thought process is shown as *trial & error* with ambiguous items, and without ambiguous items it is shown as a perceptive *insight* [15] to further look ahead in the overall process.

## 5 Conclusion

We have shown evidence that the analogy board assists creativity by forcing players to modify meanings about ambiguous items under the constrained rule.

While previous research is limited to focus on ambiguous items in terms of creativity, our findings imply that ambiguous items under constrained rule incite emotion and thought, forcing a wider reference and awareness of common functions, while analogical reasoning supports the abstraction of similarities among the bases. Based on this mind set, concept creation models could bring in repetitive *trial & error* actions that shift interpretations or *insight* behavior to look ahead for meaning of relations that depends on the existence of ambiguous items.

Further research will need to address the validation of possible patterns by semantic analysis and to deduce what words will be sufficiently ambiguous to generate new concepts.

## Acknowledgments

The work presented here was supported by the Nissan Science Foundation Grant, Japan. We thank all the baseball team members and field manager Mr.H.Nishio (teacher as well) at Kami-isshiki junior high school at Edogawa, Tokyo, Japan, for their participation to play our analogy board.

## References

1. Gentner, D., Markman, A.B.: Structure Mapping in Analogy and Similarity. American Psychologist 42(1), 45–56 (1997)
2. Waldron, R.A.: Sense and Sense Development. Andre Deutsch Ltd, London (1979)
3. Stern, G.: Meaning and Change of Meaning, Bloomington. Indiana University Press (1931)
4. Holyoak, K.J., Thagard, P.R.: Mental Leaps: Analogy in Creative Thought. MIT Press, Cambridge (1995)
5. Gomes, P., Seco, N., Pereira, F.C., Paiva, P., Carreiro, P., Ferreira, J.L., Bento, C.: The importance of retrieval in creative design analogies. Knowledge-Based Systems 19, 480–488 (2006)
6. Gentner, D., Rattermann, M.J., Forbus, K.D.: The roles of similarity in transfer: Separating retrieveability for inferential soundness. Cognitive Psychology 25, 524–575 (1993)
7. Gick, M.L., Holyoak, K.J.: Analogical problem solving. Cognitive Psychology 12, 306–355 (1980)
8. Nishihara, Y., Sunayama, W., Yachida, M.: Creative activity support by discovering effective combinations. Systems and Computers in Japan 38(12), 99–111 (2007)
9. Bellin, D., Simone, S.S.: The CRC Card Book. Addison-Wesley Pub., Reading (1997)
10. Börstler, J., Schulte, C.: Teaching Object Oriented Modelling with CRC-Cards and Role playing Games. In: Proceedings WCCE 2005 (2005)
11. Brown, S.R.: Q-methodology and Qualitative Research. Qualitative Health Research 6(4), 561–567 (1996)
12. Hori, K.: An ontology of strategic knowledge: Key concepts and applications. Knowledge-Based Systems 13(6), 369–374 (2000)
13. Bonnardel, N.: Towards understanding and supporting creativity in design: analogies in a constrained cognitive environment. Knowledge-Based Systems 13, 505–513 (2000)
14. Shank, R.C.: Dynamic memory: A theory of reminding and learning in computers and people. Cambridge University Press, Cambridge (1982)
15. Mayer, R.E.: GESTALT: Thinking as Restructuring Problems. In: Thinking, Problem solving, Cognition, ch.3, 2nd edn., pp. 39–78. W.H. Freeman and Company, New York (1992)
16. Fauconnier, G.: Mental Spaces. Cambridge University Press, Cambridge (1994)
17. Finke, R.A., Ward, T.B., Smith, S.M.: Creative Cognition. MIT Press, Cambridge (1992)
18. Finke, R.A.: Creative insight and pre-inventive forms. In: Sternberg, R.J., Davidson, J.E. (eds.) The Nature of Insight. MIT Press, Cambridge (1995)

# Exceptions as Chance for Computational Chance Discovery

Akinori Abe[1,2], Norihiro Hagita[1,3], Michiko Furutani[1], Yoshiyuki Furutani[1], and Rumiko Matsuoka[1]

[1] International Research and Educational Institute for Integrated Medical Science
(IREIIMS), Tokyo Women's Medical University
8-1 Kawada-cho, Shinjuku-ku, Tokyo 162-8666 Japan
[2] ATR Knowledge Science Laboratories
[3] ATR Intelligent Robotics and Communication Laboratories
2-2-2, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288 Japan
ave@ultimaVI.arc.net.my, hagita@atr.jp,
{michi,yoshi,rumiko}@imcir.twmu.ac.jp

**Abstract.** In this paper, we analyze clinical data to model relationships between clinical data and health levels. During analyses of data, we discovered models which are important for determining health levels but cannot be extracted during machine learning process. We regard such models as chance and propose an interactive determination of such models. The obtained models can be referred to when standard models cannot correctly explain certain individual health levels.

## 1 Introduction

In recent years, according to increased computer power, complicated data analysis on computers can be achieved. computerized data analysis could be achieved. For medical data, many data analyses are performed in various forms. Usually such analysis is conducted for data collected in usual clinical examinations or inspections. It is rather easy to collect data during such events. However, such data usually lack several parts and it is difficult to collect only perfect data sets. Therefore, in the International Research and Educational Institute for Integrated Medical Science (IREIIMS) project, we decided to collect complete medical data sets. When we collect complete data sets of considerable size, we have additional problems. The data collections include various types of data; that is, they contain data, for instance, of persons with lung cancer, those with stomach cancer etc. It is sometimes hazardous to use such mixed and complex data to perform data mining.

We have been researching chance discovery [Ohsawa and McBurney, 2003] in medical applications to deal with such complex data. In [Abe, Kogure and Hagita, 2003], for medical applications, we pointed out that it is important to find missing or hidden knowledge as a chance and showed an abductive solution. In addition, the role of induction is to determine the gray zone, which can be regarded as a chance, as well as to build a standardized knowledge base. Above we

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 750–757, 2008.

discussed the issue from a logical viewpoint, but for chance discovery, the interaction between human and computer is very important. Accordingly, Abe proposed an interactive interface for medical diagnosis support, where we can check individual data to change or confirm health levels (definition will be given in the Section 3) and discover hidden factors that might have serious or positive effects on patients health [Abe et al., 2007a]. We also proposed an integrated data Mining approach, where medical data are categorized according to given categorization (authorized by physicians) [Abe et al., 2007b] and [Abe et al., 2007a]. Since influential factors such as tumor markers control a machine learning result, a machine learning procedure can be performed for categorized data sets and results are combined or integrated. The authorized categorization is defined according to human organ situations. Therefore if we conduct machine learning for categorized data, results will be tuned according to human organ functions such as liver, pancreas, and kidney. In fact, the authorized categorization is not so small. Of course, as shown in the previous papers, machine learning methods for the authorized categorization seem to function well, but better machine learning procedures or more efficient categorization can be considered.

In this paper, we adopt the proposed interface to perform chance discovery. Especially, we focus on incorrect cases of health level determination and discuss the reasons for them. Discussions will partly be performed from the viewpoint of chance discovery. In Section 2, we briefly describe the features of the interactive data analysis system. In Section 3, we analyze (data mining) the collected data using C4.5 and apply the generated rule sets to the medical data to determine the patients' health levels. In addition, we analyze the results from the chance discovery viewpoint. Section 4 provides the final conclusion of this paper.

## 2 Medical Data Analysis System

### 2.1 Clinical Data

We are now collecting various types of clinical data, such as those obtained in blood and urine tests. In addition, health levels are assigned by doctors resulting from the clinical data and by an interview. Health levels that express the health status of patients are defined according to *Tumor stage* [Kobayashi and Kawakubo, 1994] and modified by Matsuoka. Originally, health levels consist of 5 levels (I–V). Persons at level I and II can be regarded as being healthy, but those at levels III, IV, and V can possibly develop cancer. In [Kobayashi and Kawakubo, 1994], level III is defined as the stage before the shift to preclinical cancer, level IV is defined as conventional stage 0 cancer (G0), and level V is defined as conventional stages 1–4 cancer (G1–G4). For more detailed analysis, Matsuoka defined more detailed categorization, which are I, II, III, IVa, VIb, VIc, Va, and Vb. Since health levels VI and V include many data, health levels VI and V are categorized more.

Currently, he have collected more than 2500 data with health level assignments. Medical data are mainly collected from the same client group as that

**Table 1.** Health levels

| health level | I | II | III | IVa | IVb | IVc | Va | Vb |
|---|---|---|---|---|---|---|---|---|
| ratio (%) | 0.0 | 0.0 | 3.9 | 19.5 | 52.5 | 24.1 | 7.0 | 2.6 |

described by Abe. [Abe et al., 2007b]. Thus we collected data from office workers (aged 40 to 50 years old) but not from students. Therefore, despite of a new categorization, there still exists a unbalanced health level distribution. Thus an imbalance of the data should still influence the analyses.

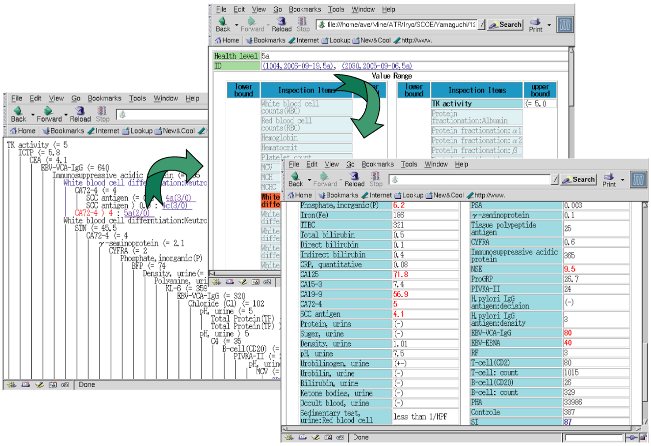
### 2.2 The Clinical Data Analysis System

We have proposed an interactive data analysis system Based on a Web browser that enables interactive checks of the learned results and corresponding data [Abe et al., 2007a],. Results generated by C4.5 [Quinlan, 1993] can be displayed on web browsers and some leaves have link anchors that can show various information on data in the decision tree. The current interactive interface is shown in Fig. 1. For instance, if we click link anchors on leaves (CA72-4 > 4: 5a(2/0); in the left browser), another browser will be started that has more detailed information (the center browser in Fig. 1). In the second (center) browser, we can check the range of each item relating to the health level. For instance, for health level 5a, the range for *TK Activity* is less than 5.0[1] In addition, we can check individual data by following links in the top of the second browser. Thus we can review results by C4.5 in the organic view and we can check each value relating to certain health levels in detail. In the following section, we use this interface to analyze clinical data in a chance discovery manner. If we use this interactive analysis, we can discover factors which do not appear in decision trees.

## 3 Chance Discovery in Medical Data Analysis

In the previous section, we illustrated the proposed interactive data analysis system. We pointed out that, by using the interface, we can check data in detail and consider background factors of situations. For medical chance discovery, we can consider various factors such as discovery of rare or novel diseases and cases which normal machine learning cannot learn. First we apply C4.5 to the collected clinical data to construct the medical knowledge base, and to reconsider the standard value of the tumor markers. Previously, we pointed out that even if induction can only generate a general knowledge set, if obtained data are in the gray zones which are outside of a standard value, the generated rules are significant as a chance [Abe, Kogure and Hagita, 2003].

[1] If the range of the data is outside of the standard value range, the item is shown in red color. The standard value is determined by referring to the data library of Mitsubishi Chemical Medicine Corporation (http://www.medience.co.jp/).



File Edit View Go Bookmarks Tools Window Help
Back Forward Reload Stop
Home Bookmarks Internet Lookup New&Cool
Done

```
TK activity <= 5
   ICTP <= 5.8
      CEA <= 4.1
         EBV-VCA-IgG <= 640
            Immunosuppressive acidic  ein <= 5
               White blood cell diffe ation:Neutro
                  CA72-4 <= 4
                     SCC antigen <= 0.    4a(3/0)
                     SCC antigen > 0.  : 4c(3/0)
                  CA72-4 > 4 : 5a(2/0)
               White blood cell differntiation:Neutro
                  STN <= 45.5
                     CA72-4 <= 4
                        γ-seminoprotein <= 2.1
                           CYFRA <= 2
                              Phosphate, inorganic(P)
                                 BFP <= 74
                                    Density, urine(=
                                       Polyamine, uri
                                          KL-6 <= 359
                                             EBV-VCA-IgG <= 320
                                                Chloride (Cl) <= 102
                                                   pH, urine <= 5
                                                      Total Protein(TP)
                                                      Total Protein(TP)
                                                   pH, urine > 5
                                                      C4 <= 35
                                                         B-cell(CD20) <=
                                                            PIVKA-II <=
                                                               pH, urine
                                                                  MCV <=
```

File Edit View Go Bookmarks Tools Window Help
Back Forward Reload Stop
file:///home/ave/Mine/ATR/Iryo/SCOE/Yamaguchi/12
Search
Print
Home Bookmarks Internet Lookup New&Cool http://www.

| Health level | 5a |
|---|---|
| ID | ⟨1004, 2006-09-19, 5a⟩, ⟨2030, 2005-09-06, 5a⟩ |

Value Range

| lower bound | Inspection Items | upper bound |
|---|---|---|
| | White blood cell counts(WBC) | |
| | Red blood cell counts(RBC) | |
| | Hemoglobin | |
| | Hematocrit | |
| | Platelet count | |
| | MCV | |
| | MCH | |
| | MCHC | |
| | White diffe | |
| | White diffe | |
| | White diffe | |
| | White diffe | |
| | White | |

| lower bound | Inspection Items | upper bound |
|---|---|---|
| | TK activity | <= 5.0 |
| | Protein fractionation:Albumin | |
| | Protein fractionation: α1 | |
| | Protein fractionation: α2 | |
| | Protein fractionation: β | |

File Edit View Go Bookmarks Tools Window Help
Back Forward Reload Stop
Search
Print
Home Bookmarks Internet Lookup New&Cool http://www.
Done

| | | | |
|---|---|---|---|
| Phosphate, inorganic(P) | 6.2 | PSA | 0.003 |
| Iron(Fe) | 186 | γ-seminoprotein | 0.1 |
| TIBC | 321 | Tissue polypeptide antigen | 25 |
| Total bilirubin | 0.5 | CYFRA | 0.6 |
| Direct bilirubin | 0.1 | Immunosuppressive acidic protein | 365 |
| Indirect bilirubin | 0.4 | NSE | 9.5 |
| CRP, quantitative | 0.08 | ProGRP | 26.7 |
| CA125 | 71.8 | PIVKA-II | 24 |
| CA15-3 | 7.4 | H.pylori IgG antigen:decision | (-) |
| CA19-9 | 56.9 | H.pylori IgG antigen:density | 3 |
| CA72-4 | 5 | EBV-VCA-IgG | 80 |
| SCC antigen | 4.1 | EBV-EBNA | 40 |
| Protein, urine | (-) | RF | 3 |
| Suger, urine | (-) | T-cell(CD2) | 80 |
| Density, urine | 1.01 | T-cell: count | 1015 |
| pH, urine | 7.5 | B-cell(CD20) | 26 |
| Urobilinogen, urine | (+-) | B-cell: count | 329 |
| Urobilin, urine | (-) | PHA | 33986 |
| Bilirubin, urine | (-) | Controle | 387 |
| Ketone bodies, urine | (-) | SI | 87 |
| Occult blood, urine | (-) | | |
| Sedimentary test, urine:Red blood cell | less than 1/HPF | | |

**Fig. 1.** Interactive interface for medical diagnosis support

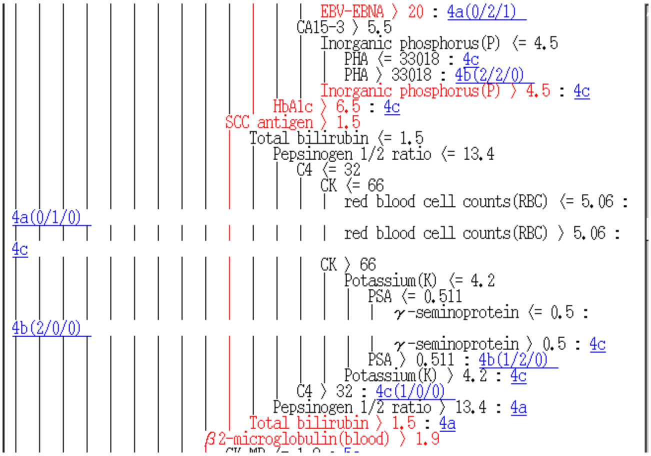
## 3.1 Health Level Determination

When we apply extracted rule sets from C4.5's result to the same data sets to determine health levels, we can obtain a rather good result, where 1403 (93.53%) of data are correctly determined, and 97 (6.67%) of data are incorrectly determined. This result is acceptable. However, when we apply the same rule sets to a collection of different data sets (963 data sets), we obtain a rather insufficient result, where 380 (39.46%) of data are correctly determined, and 583 (60.54%) of data are incorrectly determined. In fact, a part of these data is collected from the same individuals on a different date. Therefore, there will be a very small gap between learned data and health level determined data. However, a decision tree generated from 963 data sets is quite different from that generated from 1500 data sets. Even an item on the root of the decision tree is different. A reason may be that since the data set contains various types of data and the number of items is quite large (around 140), C4.5 might generate rather specialized models for the given data.

## 3.2 Health Level Determination Evaluation

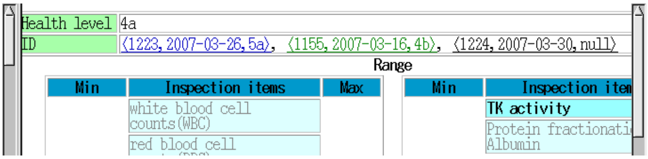
In the above, we obtained insufficient result when we apply the generated rule sets to different data set from the data set from which rule sets are generated. We adopt an interactive interface to evaluate the health level determination process.

We applied C4.5 to 1500 clinical data to obtain a decision tree (Fig.2). On the top of it (actually in the middle of the full decision tree), we can notice "EBV-EBNA $> 20$ : 4a(0/2/1)" which means two are incorrectly determined and one is assigned to a new health level. We check it because it contains two incorrect health level determinations. As shown in Fig.3, the individual whose ID is 1223 is assigned to health level 5a and the individual whose ID is 1155 is assigned to health level 4b. Health level 4b is close to 4a, so it will be acceptable. In fact, the



```
                    EBV-EBNA > 20 : 4a(0/2/1)
                  CA15-3 > 5.5
                    Inorganic phosphorus(P) <= 4.5
                      PHA <= 33018 : 4c
                      PHA > 33018 : 4b(2/2/0)
                    Inorganic phosphorus(P) > 4.5 : 4c
                HbA1c > 6.5 : 4c
              SCC antigen > 1.5
                Total bilirubin <= 1.5
                  Pepsinogen 1/2 ratio <= 13.4
                    C4 <= 32
                      CK <= 66
                        red blood cell counts(RBC) <= 5.06 :
4a(0/1/0)
                        red blood cell counts(RBC) > 5.06 :
4c
                      CK > 66
                        Potassium(K) <= 4.2
                          PSA <= 0.511
                            γ-seminoprotein <= 0.5 :
4b(2/0/0)
                            γ-seminoprotein > 0.5 : 4c
                          PSA > 0.511 : 4b(1/2/0)
                        Potassium(K) > 4.2 : 4c
                    C4 > 32 : 4c(1/0/0)
                  Pepsinogen 1/2 ratio > 13.4 : 4a
                Total bilirubin > 1.5 : 4a
              β2-microglobulin(blood) > 1.9
```

**Fig. 2.** Part of decision tree



| Health level | 4a |
|---|---|
| ID | <1223, 2007-03-26, 5a>, <1155, 2007-03-16, 4b>, <1224, 2007-03-30, null> |

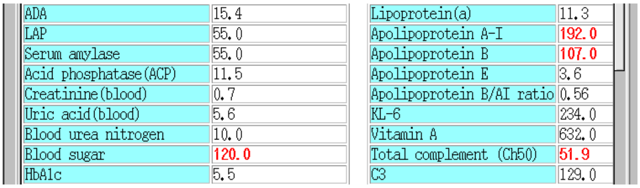
Range

| Min | Inspection items | Max | Min | Inspection iten |
|---|---|---|---|---|
| | white blood cell counts(WBC) | | | TK activity |
| | red blood cell | | | Protein fractionati Albumin |

**Fig. 3.** Part of second browser

patient had rather abnormal values in several items such as NSE (6.2), EBV-VCA-IgG (160.0), and EBV-EBNA (40.0). However, as for 5a, there should be certain problems in determining health levels. Therefore, we review the client (ID: 1223)'s data in detail.

For the data review, we utilized the interactive interface. When we click the patient's ID (in this case, 1223) in the second browser, we can obtain the third browser shown in Fig.4, which presents patient's clinical data. In the browser, we can easily find the abnormally high values written in red (in BW print, light gray). For instance, Apolipoprotein A-I is 192.0, Apolipoprotein B is 107.0, and blood sugar is 120. That means, the patient suffers from rather serious Lipid metabolism abnormalities or Arteriosclerosis. In fact, with the proposed interface, it is rather easy to focus on such abnormal data. Why did rules extracted from C4.5 determine the patient's health level as 4a instead of 5a? When the rule set was applied to the patient data, neither Apolipoprotein A-I, Apolipoprotein B nor blood sugar was used to determine his/her health level. Apolipoprotein B is not in the path after "TK activity <= 5.4." His/her TK activity is 5.4, but this is not a serious problem. Even if we follow the path after "TK activity > 5.4," neither of the above factors is used to determine his/her health level. In fact, Apolipoprotein B appears in the decision tree. Accordingly, we apply the rule set that was only generated from metabolic function test data including



| | | | |
|---|---|---|---|
| ADA | 15.4 | Lipoprotein(a) | 11.3 |
| LAP | 55.0 | Apolipoprotein A-I | 192.0 |
| Serum amylase | 55.0 | Apolipoprotein B | 107.0 |
| Acid phosphatase(ACP) | 11.5 | Apolipoprotein E | 3.6 |
| Creatinine(blood) | 0.7 | Apolipoprotein B/AI ratio | 0.56 |
| Uric acid(blood) | 5.6 | KL-6 | 234.0 |
| Blood urea nitrogen | 10.0 | Vitamin A | 632.0 |
| Blood sugar | 120.0 | Total complement (Ch50) | 51.9 |
| HbA1c | 5.5 | C3 | 129.0 |

**Fig. 4.** Part of a third browser

Apolipoprotein A-I, Apolipoprotein B, and blood sugar. In this case, the health level of a patient (ID: 1223) is 4b. Due to an imbalance influential power of attributes, sometimes generated rules (models) are influenced by such powerful factors. In [Abe et al., 2007b], we proposed an integrated data mining to remove influential factors which might disturb a proper model generation. Thus, in this application, the influence of TK activity (tumor marker) should be removed in this process, but still we cannot make a correct health level determination.

We combine the data collection with 1500 and 963 data sets to apply C4.5. We obtained quite different results from those that were generated separately from both data sets. The result shows that very different patient groups will be added to the original data sets (1500 data set). In fact, at the top of the decision tree, albumin which is categorized for liver, pancreas, and kidney test data appears. A patient (ID: 1223) is in the group whose feature seems to be the same as or similar to the original data. However, based on the result above, he/she might belong to the other data group. The main factors which determine the patient (ID: 1223)'s health level will belong to metabolic function test data, which cannot be considered in the health level determination process.

### 3.3 Model Generation and Chance Discovery

As shown above, from the current data set, it will be difficult to model relationships between clinical data and health levels only by applying C4.5. In fact, it is necessary to collect more data to model relationships between clinical data and health levels. In addition, we can point out that one of the serious weak points of the decision tree approach is that we cannot make multiple models in a simple way. That is, the top of decision tree is one model. For the other analysis, we apply Clementine to perform basket analysis [Agrawal et al., 1993]. We conducted a 3-item-sets basket analysis[2]. For the health level 5a, we obtained that the set consisting of "Immunosuppressive acidic protein and ALP Type 2 isoenzyme," "Immunosuppressive acidic protein and Apolipoprotein E, " "Alpha1-Globulin and Apolipoprotein E, " etc. are factors for determining health levels. Thus we can also obtain categorization information from the result. Thus we can model relationships between clinical data and health levels by applying both C4.5 and basket analysis, where basket analysis can support categorized modeling for C4.5.

[2] For reasons of computational limitation, we could not perform more than 4-item-sets basket analyses.

However, as shown above, there still exist exceptions such as the case of patient (ID: 1223). For chance discovery modelling, it would be necessary to analyze such exceptional cases. And it would be necessary to generate missing or hidden paths or new rules to explain events that cannot be explained by standard models. For such explanations, usually abduction can be applied. If a part of a decision tree contains necessary rule clusters, abduction can suggest a necessary path to the clusters. In the above case, we should build a new model to explain the situation. That is, if models cannot properly determine health levels, a new model that can explain an exception should be added to the current model. After that, we can consider additional models during health level determination by the standard model generated for instance with C4.5. We can regard such additional models as chance in the context of chance discovery which do not frequently appear but are very significant for determination of exceptional cases.

In fact, these types of exceptions are not regarded as exception when we check data personally, because we can focus on abnormal data easily. In addition, for medical diagnosis, such cases are reported as abnormal case for the health. However, if computers view data based on the own generalized models to determine health levels, it is not easy to focus on such exceptions. From a computational viewpoint, the preparation of such exceptional models is very important to perform a proper health level determination. In the case, standard models can be given up and exceptional models can be applied to determine health levels. That is, a computational procedure should jump to rare or novel models when necessary.

## 4 Conclusions

In this paper, we discuss computational generation of relationships between health levels and clinical data. For model generation, we mainly adopt C4.5 which can generate a decision tree. In fact, C4.5 is usually used for general model generation, but we discuss such model generation from the chance discovery viewpoint, because we cannot generate proper models which can explain most of a data set. In fact, the generated model can properly explain only data from which the model is generated. After analyses of incorrect explanations (health level determinations), we addressed the importance of rule sets which cannot be extracted by standard machine learning methods. Such sets are important to determine health levels but since some events do not frequently appear, it is difficult to model relationships between such events and health levels. We proposed an interactive discovery of such events by using the proposed web-based interface.

For discovered rare rule sets, we pointed out that it is necessary to generate a missing or hidden paths or new rule to reach rare or novel models. We also point out the possibility of adoption of abduction to generate missing or hidden paths. In this paper, we did not propose a concrete strategy to generate missing or hidden paths or new rules. In the future it is necessary to propose a strategy to suggest missing or hidden paths or new rules. An abductive discovery approach will be one of the solutions.

## Acknowledgments

This research was supported in part by the Program for Promoting the Establishment of Strategic Research Centers, Special Coordination Funds for Promoting Science and Technology, Ministry of Education, Culture, Sports, Science and Technology (Japan).

## References

[Abe, Kogure and Hagita, 2003] Abe, A., Kogure, K., Hagita, N.: Discovery of Hidden Relations from Medical Data. In: Proc. of HCI 2003 3rd. Int'l Workshop on Chance Discovery, pp. 37–43 (2003)

[Abe et al., 2007a] Abe, A., Hagita, N., Furutani, M., Furutani, Y., Matsuoka, R.: An interface for medical diagnosis support. In: Apolloni, B., Howlett, R.J., Jain, L. (eds.) KES 2007, Part II. LNCS (LNAI), vol. 4693, pp. 909–916. Springer, Heidelberg (2007)

[Abe et al., 2007b] Abe, A., Hagita, N., Furutani, M., Furutani, Y., Matsuoka, R.: Possibility of Integrated Data Mining of Clinical Data. Data Science Journal 6(suppl.), S104–S115 (2007)

[Abe et al., 2007c] Abe, A., Hagita, N., Furutani, M., Furutani, Y., Matsuoka, R.: Data mining of Multi-categorized Data. In: Raś, Z.W., Tsumoto, S., Zighed, D.A. (eds.) MCD 2007. LNCS (LNAI), vol. 4944, pp. 209–220. Springer, Heidelberg (2008)

[Agrawal et al., 1993] Agrawal, R., Imielinski, T., Swami, A.: Mining association rules between sets of items in large databases. In: Proc. of ACM SIGMOD Int'l Conf. on Management of Data, pp. 207–216 (1993)

[Kobayashi and Kawakubo, 1994] Kobayashi, T., Kawakubo, T.: Prospective Investigation of Tumor Markers and Risk Assessment in Early Cancer Screening. Cancer 73(7), 1946–1953 (1994)

[Ohsawa and McBurney, 2003] Osawa, Y., McBurney, P. (eds.): Chance Discovery. Springer, Heidelberg (2003)

[Quinlan, 1993] Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufman, San Francisco (1993)

# Analyzing Behavior of Objective Rule Evaluation Indices Based on a Correlation Coefficient

Hidenao Abe and Shusaku Tsumoto

Shimane University
89-1 Enya-cho Izumo Shimane, 6938501, Japan
abe@med.shimane-u.ac.jp, tsumoto@computer.org

**Abstract.** In this paper, we present an analysis of behavior of objective rule evaluation indices on classification rule sets using Pearson product-moment correlation coefficients. To support data mining post-processing, which is one of important procedures in a data mining process, at least 40 indices are proposed to find out valuable knowledge. However, their behavior have never been clearly articulated. Therefore, we carried out a correlation analysis between each objective rule evaluation index. In this analysis, we calculated average values of each index using bootstrap method on 32 classification rule sets learned with information gain ratio. Then, we found the following relationships based on the correlation coefficient values: similar pairs, discrepant pairs, and independent indices. With regarding to this result, we discuss about relative functional relationships between each group of objective indices.

## 1 Introduction

In recent years, enormous amounts of data have been stored on information systems in natural science, social science, and business domains. People have been able to obtain valuable knowledge due to the development of information technology. Besides, data mining techniques combine different kinds of technologies such as database technologies, statistical methods, and machine learning methods. Then, data mining has been well known for utilizing data stored on database systems. In particular, if-then rules, which are produced by rule induction algorithms, are considered as one of the highly usable and readable outputs of data mining. However, to large datasets with hundreds of attributes including noise, the process often obtains many thousands of rules. From such a large rule set, it is difficult for human experts to find out valuable knowledge, which are rarely included in the rule set.

To support such a rule selection, many studies have done using objective rule evaluation indices such as recall, precision, and other interestingness measurements [1,2,3] (Hereafter, we refer to these indices as "objective indices"). Although their properties are identified with their definitions, their behavior on rule sets are not investigated with any promising method.

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 758–765, 2008.

With regard to the above-mentioned issues, we present a correlation analysis method to identify the functional properties of objective indices in Section 3. Then, with the 39 objective indices and classification rule sets from 32 UCI datasets, we identified the following relationships based on the correlation analysis method: similar pairs of indices, contradict pairs of indices, and independent indices. Based on the result in Section 4, we discuss about these relationships and differences between functional properties and original definitions.

## 2 Interestingness Measures and Related Work

Many studies have investigated the selection of valuable rules from a large mined rule set based on objective rule evaluation indices. Some of these works suggested the indices to discover interesting rules from such a large number of rules [1,2,3]. These interestingness measures are based on two different approaches[4]: the objective (data-driven) approach and the subjective approach.

To avoid confusing real human interest, the objective index, and the subjective index, we clearly define these three items as follows: **Objective Index :** features such as the correctness, uniqueness, and strength of a rule, which are calculated mathematically. An objective index does not include any human evaluation criteria. **Subjective Index :** The similarity or difference between the information on interestingness given beforehand by a human expert and that obtained from a rule. Although some human criteria are included in its initial state, the similarity or difference is mainly calculated mathematically.

However, there has been not yet done to analyze some functional relationships among objective indices on any actually obtained classification rule set totally.

## 3 Correlation Analysis for the Objective Rule Evaluation Indices

In this section, we describe a correlation analysis method to identify behavior of objective indices. To analyze functional relationships between objective indices, we should gather the following materials: values of objective indices of each classification rule set learned from each dataset, and correlation coefficients between objective indices with the values. The process of the analysis is shown in Figure 1.

First, we obtain multiple rule sets from some datasets to get values of objective indices. When gathering these values, we should care the statistical correctness of each value. Therefore, the values are averaged adequately large number ($> 100$) of values from bootstrap samples.

Then, Pearson product-moment correlation coefficients $r$ between indices, $x$ and $y$, are calculated for $n$ datasets.

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

With these coefficient values, we identified similar pairs, contradict pairs, and independent indices.

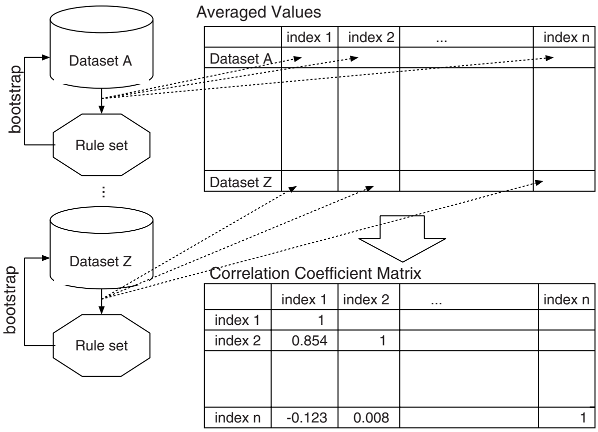


**Fig. 1.** An overview of the correlation analysis method

## 4 Analyzing the Objective Rule Evaluation Indices on UCI Datasets

In this section, we describe the correlation analysis of the 39 objective indices with 32 UCI datasets. Table 1 shows the 39 objective indices investigated and reformulated for classification rules by Ohsaki et. al.[5].

As for datasets, we have taken the 32 datasets from UCI machine learning repository[18], which are distributed with Weka [19].

For the above datasets, we obtained rule sets with PART [20] implemented in Weka. PART constructs a rule set based on information gain ratio. This means the obtained rule sets are biased with the correctness of classification.

### 4.1 Constructing a Correlation Coefficient Matrix of the 39 Objective Indices

For the 32 datasets, we obtained the rule sets using PART. This procedure is repeated 1000 times with bootstrap re-sampling for each dataset. As a representative value for each bootstrap iteration, the average for a rule set has been calculated. Then, we averaged the average values from 1000 times iterations.

With the average values for each dataset, we calculated correlation coefficients between each objective index.

### 4.2 Identifying Characteristic Relationships between Objective Indices Based on Correlation Coefficient Matrix Analysis

Based on the correlation coefficients, we identify characteristic relationship between each objective index. We defined the three characteristic relation ship as follows:

- Similar pair: two indices has strong positive correlation $r > 0.8$.
- Discrepant pair: two indices has strong negative correlation $r < -0.8$.
- Independent index: a index has only weak correlations $-0.8 \leq r \leq 0.8$ for the other indices.

Figure 2 shows an overview of similar and discrepant pairs of the objective indices on the correlation analysis. There are several groups having mutual correlations.

Table 2 shows similar pairs of objective indices on the correlation analysis. There are several groups having mutual correlations. The largest group, which has correlation to Cosine Similarity and F-Measure, includes 23 indices. Relative Risk and Odds Ratio make another group. $\chi^2$-M1, $\chi^2$-M4 and PSI also make different functional group. These pairs indicate distinct functional property for the rule sets.

As shown in Table 3, there are smaller number of discrepant pairs of the objective indices on the correlation analysis. Accuracy and Prevalence have discrepant behavior each other. Likewise, BI and BC also indicate discrepant behavior based on the negative correlation between them. BC shows discrepant behavior to several indices, which belong to the biggest group of the similar pairs.

Figure 3 shows scatter plots of representative pair of each relationship. Where $r = 1.00$ is not correctly 1. Also, $r = -1.00$ is not correctly $-1$.

**Table 1.** Objective rule evaluation indices for classification rules used in this research. **P:** Probability of the antecedent and/or consequent of a rule. **S:** Statistical variable based on P. **I:** Information of the antecedent and/or consequent of a rule. **N:** Number of instances included in the antecedent and/or consequent of a rule. **D:** Distance of a rule from the others based on rule attributes.

| **Theory** | Index Name (**Abbreviation**) [Reference Number of Literature] |
|---|---|
| **P** | Coverage (**Coverage**), Prevalence (**Prevalence**)<br>Precision (**Precision**), Recall (**Recall**)<br>Support (**Support**), Specificity (**Specificity**)<br>Accuracy (**Accuracy**), Lift (**Lift**)<br>Leverage (**Leverage**), Added Value (**Added Value**)[2]<br>Klösgen's Interestingness (**KI**)[6], Relative Risk (**RR**)[7]<br>Brin's Interest (**BI**)[8], Brin's Conviction (**BC**)[8]<br>Certainty Factor (**CF**)[2], Jaccard Coefficient (**Jaccard**)[2]<br>F-Measure (**F-M**)[9], Odds Ratio (**OR**)[2]<br>Yule's Q (**YuleQ**)[2], Yule's Y (**YuleY**)[2]<br>Kappa (**Kappa**)[2], Collective Strength (**CST**)[2]<br>Gray and Orlowska's Interestingness weighting Dependency (**GOI**)[10]<br>Gini Gain (**Gini**)[2], Credibility (**Credibility**)[11] |
| **S** | $\chi^2$ Measure for One Quadrant ($\chi^2$**-M1**)[12]<br>$\chi^2$ Measure for Four Quadrant ($\chi^2$**-M4**)[12] |
| **I** | J-Measure (**J-M**)[13], K-Measure (**K-M**)[14]<br>Mutual Information (**MI**)[2]<br>Yao and Liu's Interestingness 1 based on one-way support (**YLI1**)[3]<br>Yao and Liu's Interestingness 2 based on two-way support (**YLI2**)[3]<br>Yao and Zhong's Interestingness (**YZI**)[3] |
| **N** | Cosine Similarity (**CSI**)[2], Laplace Correction (**LC**)[2]<br>$\phi$ Coefficient ($\phi$)[2], Piatetsky-Shapiro's Interestingness (**PSI**)[15] |
| **D** | Gago and Bento's Interestingness (**GBI**)[16]<br>Peculiarity (**Peculiarity**)[17] |

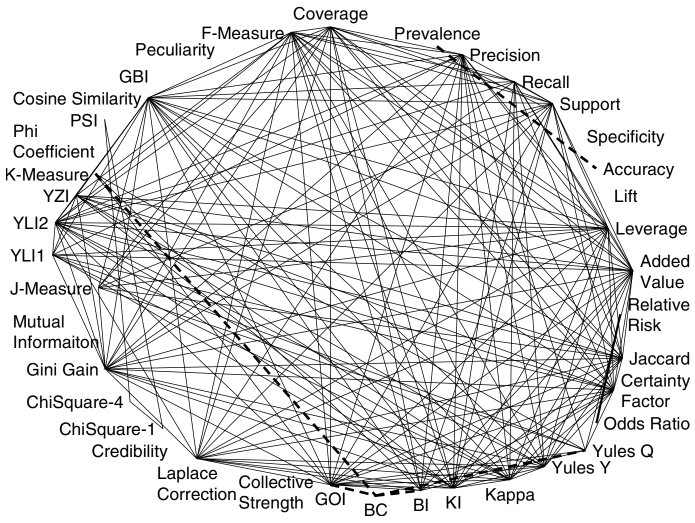


**Fig. 2.** Similar pairs (with solid line) and discrepant pairs (with dotted line) of the objective indices on the correlation analysis

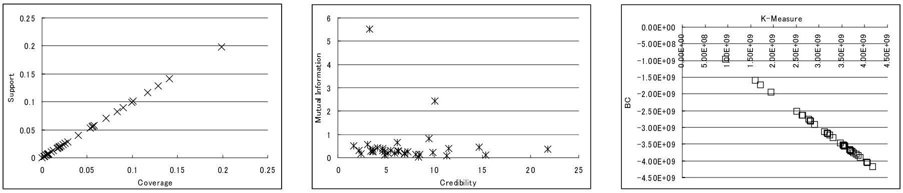


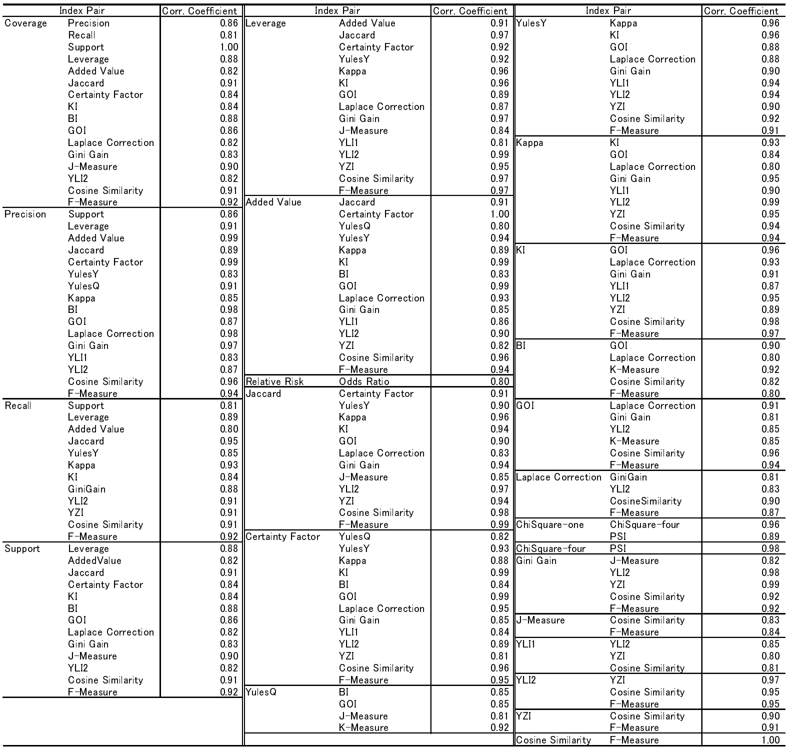
**Fig. 3.** Scatter plots of representative pairs. (Coverage vs. Support ($r = 1.00$), Credibility vs. Mutual Information ($r = -0.10$) and K-Measure vs. BC ($r = -1.00$)).

## 5 Discussion

With regarding to Figure 2 and Table 2, we can say that the following objective indices indicate similar behavior: Coverage, Precision, Recall, Support, Leverage, Added Value, Jaccard, Certainty Factor, YulesQ, YulesY, Kappa, KI, BI, GOI, Laplace Correction, Gini Gain, J-Measure, YLI1, YLI2, YZI, K-Measure, Cosine Similarity, and F-Measure. The other groups also show similar behavior to the classification rule sets based on information gain ratio. Considering their definitions, although they have different theoretical backgrounds, their functional property is to represent correctness of rules. This indicates that these indices evaluate given rules optimistically.

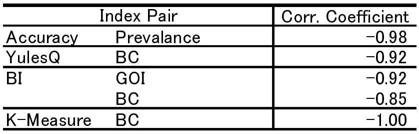
On the other hand, BC indicates opposite behavior comparing with the former indices. Therefore, the result indicates that BC evaluate given rules not so optimistically. As for Accuracy and Prevalence, Accuracy measures ratio of both

**Table 2.** Similar pairs of objective indices on the correlation analysis



| Index Pair | | Corr. Coefficient |
|---|---|---|
| Coverage | Precision | 0.86 |
| | Recall | 0.81 |
| | Support | 1.00 |
| | Leverage | 0.88 |
| | Added Value | 0.82 |
| | Jaccard | 0.91 |
| | Certainty Factor | 0.84 |
| | KI | 0.84 |
| | BI | 0.88 |
| | GOI | 0.86 |
| | Laplace Correction | 0.82 |
| | Gini Gain | 0.83 |
| | J-Measure | 0.90 |
| | YLI2 | 0.82 |
| | Cosine Similarity | 0.91 |
| | F-Measure | 0.92 |
| Precision | Support | 0.86 |
| | Leverage | 0.91 |
| | Added Value | 0.99 |
| | Jaccard | 0.89 |
| | Certainty Factor | 0.99 |
| | YulesY | 0.83 |
| | YulesQ | 0.91 |
| | Kappa | 0.85 |
| | BI | 0.98 |
| | GOI | 0.87 |
| | Laplace Correction | 0.98 |
| | Gini Gain | 0.97 |
| | YLI1 | 0.83 |
| | YLI2 | 0.87 |
| | Cosine Similarity | 0.96 |
| | F-Measure | 0.94 |
| Recall | Support | 0.81 |
| | Leverage | 0.89 |
| | Added Value | 0.80 |
| | Jaccard | 0.95 |
| | YulesY | 0.85 |
| | Kappa | 0.93 |
| | KI | 0.84 |
| | GiniGain | 0.88 |
| | YLI2 | 0.91 |
| | YZI | 0.91 |
| | Cosine Similarity | 0.91 |
| | F-Measure | 0.92 |
| Support | Leverage | 0.88 |
| | AddedValue | 0.82 |
| | Jaccard | 0.91 |
| | Certainty Factor | 0.84 |
| | KI | 0.84 |
| | BI | 0.88 |
| | GOI | 0.86 |
| | Laplace Correction | 0.82 |
| | Gini Gain | 0.83 |
| | J-Measure | 0.90 |
| | YLI2 | 0.82 |
| | Cosine Similarity | 0.91 |
| | F-Measure | 0.92 |
| Leverage | Added Value | 0.91 |
| | Jaccard | 0.97 |
| | Certainty Factor | 0.92 |
| | YulesY | 0.92 |
| | Kappa | 0.96 |
| | KI | 0.96 |
| | GOI | 0.89 |
| | Laplace Correction | 0.87 |
| | Gini Gain | 0.97 |
| | J-Measure | 0.84 |
| | YLI1 | 0.81 |
| | YLI2 | 0.99 |
| | YZI | 0.95 |
| | Cosine Similarity | 0.97 |
| | F-Measure | 0.97 |
| Added Value | Jaccard | 0.91 |
| | Certainty Factor | 1.00 |
| | YulesQ | 0.80 |
| | YulesY | 0.94 |
| | Kappa | 0.89 |
| | KI | 0.99 |
| | BI | 0.83 |
| | GOI | 0.99 |
| | Laplace Correction | 0.93 |
| | Gini Gain | 0.85 |
| | YLI1 | 0.86 |
| | YLI2 | 0.90 |
| | YZI | 0.82 |
| | Cosine Similarity | 0.96 |
| | F-Measure | 0.94 |
| Relative Risk | Odds Ratio | 0.80 |
| Jaccard | Certainty Factor | 0.91 |
| | YulesY | 0.90 |
| | Kappa | 0.96 |
| | KI | 0.94 |
| | GOI | 0.90 |
| | Laplace Correction | 0.83 |
| | Gini Gain | 0.94 |
| | J-Measure | 0.85 |
| | YLI2 | 0.97 |
| | YZI | 0.94 |
| | Cosine Similarity | 0.98 |
| | F-Measure | 0.99 |
| Certainty Factor | YulesQ | 0.82 |
| | YulesY | 0.93 |
| | Kappa | 0.88 |
| | KI | 0.99 |
| | BI | 0.84 |
| | GOI | 0.99 |
| | Laplace Correction | 0.95 |
| | Gini Gain | 0.85 |
| | YLI1 | 0.84 |
| | YLI2 | 0.89 |
| | YZI | 0.81 |
| | Cosine Similarity | 0.96 |
| | F-Measure | 0.95 |
| YulesQ | BI | 0.85 |
| | GOI | 0.85 |
| | J-Measure | 0.81 |
| | K-Measure | 0.92 |
| YulesY | Kappa | 0.96 |
| | KI | 0.96 |
| | GOI | 0.88 |
| | Laplace Correction | 0.88 |
| | Gini Gain | 0.90 |
| | YLI1 | 0.94 |
| | YLI2 | 0.94 |
| | YZI | 0.90 |
| | Cosine Similarity | 0.92 |
| | F-Measure | 0.91 |
| Kappa | KI | 0.93 |
| | GOI | 0.84 |
| | Laplace Correction | 0.80 |
| | Gini Gain | 0.95 |
| | YLI1 | 0.90 |
| | YLI2 | 0.99 |
| | YZI | 0.95 |
| | Cosine Similarity | 0.94 |
| | F-Measure | 0.94 |
| KI | GOI | 0.96 |
| | Laplace Correction | 0.93 |
| | Gini Gain | 0.91 |
| | YLI1 | 0.87 |
| | YLI2 | 0.95 |
| | YZI | 0.89 |
| | Cosine Similarity | 0.98 |
| | F-Measure | 0.97 |
| BI | GOI | 0.90 |
| | Laplace Correction | 0.80 |
| | K-Measure | 0.92 |
| | Cosine Similarity | 0.82 |
| | F-Measure | 0.80 |
| GOI | Laplace Correction | 0.91 |
| | Gini Gain | 0.81 |
| | YLI2 | 0.85 |
| | K-Measure | 0.85 |
| | Cosine Similarity | 0.96 |
| | F-Measure | 0.94 |
| Laplace Correction | GiniGain | 0.81 |
| | YLI2 | 0.83 |
| | CosineSimilarity | 0.90 |
| | F-Measure | 0.87 |
| ChiSquare-one | ChiSquare-four | 0.96 |
| | PSI | 0.89 |
| ChiSquare-four | PSI | 0.98 |
| Gini Gain | J-Measure | 0.82 |
| | YLI2 | 0.98 |
| | YZI | 0.99 |
| | Cosine Similarity | 0.92 |
| | F-Measure | 0.92 |
| J-Measure | Cosine Similarity | 0.83 |
| | F-Measure | 0.84 |
| YLI1 | YLI2 | 0.85 |
| | YZI | 0.80 |
| | Cosine Similarity | 0.81 |
| YLI2 | YZI | 0.97 |
| | Cosine Similarity | 0.95 |
| | F-Measure | 0.95 |
| YZI | Cosine Similarity | 0.90 |
| | F-Measure | 0.91 |
| Cosine Similarity | F-Measure | 1.00 |

**Table 3.** Discrepant pairs of objective indices on the correlation analysis



| Index Pair | | Corr. Coefficient |
|---|---|---|
| Accuracy | Prevalance | −0.98 |
| YulesQ | BC | −0.92 |
| BI | GOI | −0.92 |
| | BC | −0.85 |
| K-Measure | BC | −1.00 |

of true positive and false negative for each rule. On the other hand, Prevalence only measures ratio of mentioned class value of each rule. It is reasonable to indicate discrepant property, because accurate rules have high Accuracy values irrespective of their mentioned class value.

As for the independent indices, GBI and Peculiarity suggested with the different theoretical background comparing with the other indices. Therefore, what they have different behavior is reasonable. However, Corrective Strength, Credibility, Mutual Information and $\phi$ Coefficient indicate the different behavior, comparing with the other indices which have the same theoretical backgrounds (**P**,**S** and **N**). These indices evaluate given rules from each different viewpoint.

## 6 Conclusion

In this paper, we described the method to analyze functional properties of objective rule evaluation indices.

We investigated functional properties of objective indices with 32 UCI datasets and their rule sets as an actual example. With regarding to the result, several groups are found as functional similarity groups in cross-sectional manner for their theoretical backgrounds.

In the future, we will investigate functional properties of objective indices to other kind of rule sets obtained from the other rule mining algorithms. At the same time, we will investigate not only Pearson product-moment correlation coefficient but also rank correlation coefficients and other correlations.

## References

1. Hilderman, R.J., Hamilton, H.J.: Knowledge Discovery and Measure of Interest. Kluwer Academic Publishers, Dordrecht (2001)
2. Tan, P.N., Kumar, V., Srivastava, J.: Selecting the right interestingness measure for association patterns. In: Proceedings of International Conference on Knowledge Discovery and Data Mining KDD 2002, pp. 32–41 (2002)
3. Yao, Y.Y., Zhong, N.: An analysis of quantitative measures associated with rules. In: Zhong, N., Zhou, L. (eds.) PAKDD 1999. LNCS (LNAI), vol. 1574, pp. 479–488. Springer, Heidelberg (1999)
4. Freitas, A.A.: On rule interestingness measures. Knowledge-Based Systtems 12(5-6), 309–315 (1999)
5. Ohsaki, M., Abe, H., Yokoi, H., Tsumoto, S., Yamaguchi, T.: Evaluation of rule interestingness measures in medical knowledge discovery in databases. Artificial Intelligence in Medicine 41(3), 177–196 (2007)
6. Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (eds.) : Explora: A Multipattern and Multistrategy Discovery Assistant. In: Advances in Knowledge Discovery and Data Mining, pp. 249–271. AAAI/MIT Press, California (1996)
7. Ali, K., Manganaris, S., Srikant, R.: Partial classification using association rules. In: Proceedings of the International Conference on Knowledge Discovery and Data Mining KDD 1997, pp. 115–118 (1997)
8. Brin, S., Motwani, R., Ullman, J., Tsur, S.: Dynamic itemset counting and implication rules for market basket data. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, pp. 255–264 (1997)
9. Rijsbergen, C.: Information retrieval, ch. 7 (1979), http://www.dcs.gla.ac.uk/Keith/Chapter.7/Ch.7.html
10. Gray, B., Orlowska, M.E.: CCAIIA: Clustering categorical attributes into interesting association rules. In: Wu, X., Kotagiri, R., Korb, K.B. (eds.) PAKDD 1998. LNCS, vol. 1394, pp. 132–143. Springer, Heidelberg (1998)
11. Hamilton, H.J., Shan, N., Ziarko, W.: Machine learning of credible classifications. In: Australian Conf. on Artificial Intelligence AI 1997, pp. 330–339 (1997)
12. Goodman, L.A., Kruskal, W.H.: Measures of association for cross classifications. Springer Series in Statistics, vol. 1. Springer, Heidelberg (1979)
13. Smyth, P., Goodman, R.M.: Rule induction using information theory. In: Piatetsky-Shapiro, G., Frawley, W.J. (eds.) Knowledge Discovery in Databases, pp. 159–176. AAAI/MIT Press (1991)

14. Ohsaki, M., Kitaguchi, S., Kume, S., Yokoi, H., Yamaguchi, T.: Evaluation of rule interestingness measures with a clinical dataset on hepatitis. In: Boulicaut, J.-F., Esposito, F., Giannotti, F., Pedreschi, D. (eds.) PKDD 2004. LNCS (LNAI), vol. 3202, pp. 362–373. Springer, Heidelberg (2004)
15. Piatetsky-Shapiro, G.: Discovery, analysis and presentation of strong rules. In: Piatetsky-Shapiro, G., Frawley, W.J. (eds.) Knowledge Discovery in Databases, pp. 229–248. AAAI/MIT Press (1991)
16. Gago, P., Bento, C.: A metric for selection of the most promising rules. In: European Conference on the Principles of Data Mining and Knowledge Discovery PKDD 1998, pp. 19–27 (1998)
17. Zhong, N., Yao, Y.Y., Ohshima, M.: Peculiarity oriented multi-database mining. IEEE Transactions on Knowledge and Data Engineering 15(4), 952–960 (2003)
18. Hettich, S., Blake, C.L., Merz, C.J.: UCI repository of machine learning databases. Department of Information and Computer Science, University of California, Irvine (1998), http://www.ics.uci.edu/~mlearn/MLRepository.html
19. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, San Francisco (2000)
20. Frank, E., Witten, I.H.: Generating accurate rule sets without global optimization. In: The Fifteenth International Conference on Machine Learning, pp. 144–151 (1998)

# Face Image Annotation Based on Latent Semantic Space and Rules

Hideaki Ito, Yuji Kawai, and Hiroyasu Koshimizu

School of Information Science and Technology, Chukyo University
101 Tokodachi, Kaizu-cho, Toyota, Aichi, 470-0393 Japan
{itoh@sist, h104045@st, hiroyasu@sist}@chukyo-u.ac.jp

**Abstract.** This paper presents a face image annotation system based on latent semantic indexing and rules. To achieve annotation, visual and symbolic features are integrated. Two features are corresponding to lengths and/or widths of face parts and keywords, respectively. In order to develop annotation mechanism, it is required to vary the dimensions of the spaces which are constructed by the latent semantic indexing, and to represent direct relationships among features. Associated symbolic features to visual features are represented in rules based on decision trees. Co-occurrence relationships among keywords are represented in association rules.

**Keywords:** face image annotation, latent semantic indexing, decision tree, association rule.

## 1 Introduction

In recent years, image data are accumulated enormously, it is necessary to develop image annotation systems [2,8]. Usually, objects or regions of natural images are described in words. However, in face image annotation systems, it is desired to represent inspired impressions from face images [4].

We have been developing a face image annotation system, named FIARS [6]. The purposes to develop this system are to retrieve face images in keywords, and to assign keywords to face images. Keywords are selected after emphasizing characters on faces. The characters are depicted by comparing one person and other persons. For example, when a caricature of a person is drawn, emphasis are made by appropriately modifying measured characters [1]. And, it is considered that the emphasized characters on the face are represented in terms of keywords, such as, round eye, thin lip, large nose, etc.

Annotation is achieved by integrating visual and symbolic features. Visual features are lengths and/or widths of individual face parts, called part data. Symbolic features are keywords which describe impressions with respect to sizes and/or shapes of a face. These features are integrated by latent semantic indexing [11] in FIARS. However, during progress of this system, it is required to specify associations from visual features to symbolic features in direct, and to adjust dimensions of constructed latent semantic spaces. For meeting them,

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 766–773, 2008.

two mechanisms are developed; one is to construct rules, and another is to treat arbitrary dimensions of the spaces. Decision trees and association rules [5] are constructed. Decision trees specify direct relationships between part data and keywords. Association rules specify associations among keywords. On the other hand, retrieval results are changed according to the dimensions of the spaces.

Many annotation systems are developed by integrating two types of features. Rules to specify relationships between visual and symbolic features are proposed by [3]. Textures and colors are used as visual features. An annotation system based on latent semantic indexing is developed [13]. [9] utilizes probabilistic latent semantic indexing. On the other hand, many systems are developed for recognizing faces [1]. Some points or regions on a face are measured to meet inherent applications. In FIARS, lengths/widths and distances of face parts are measured, since characters are represented in keywords with respect to sizes or shapes. Moreover, an automatic facial expression analysis system is developed for human emotion analysis using face action units [10]. They are used to represent emotions, such as happiness, anger, fear, etc. FIARS deals with keywords which describe characters of face parts, such as thin lip, slender eyebrow, small eye, round face, etc.

This paper is organized as follows. Section 2 shows an overview of the system. Description of face images is shown in Sec. 3. Section 4 describes annotation mechanisms based on latent semantic indexing, decision trees, association rules, and their efficiencies. Finally, concluding remarks are described in Sec. 5.

## 2 An Overview

During the development of FIARS [6], it is desired to decide suitable dimensions of constructed spaces, and to describe relations among part data and keywords. The following three mechanisms are developed for meeting them. They are; to specify dimensions of constructed spaces; to make decision rules which specify relationships between part data and keywords, constructed from decision trees; and to make association rules which specify co-occurrence of keywords.

Figure 1 shows an overview of the system. When a face image is given, a set of keywords is assigned finally. To achieve this, a given face image is compared with existing faces. A set of face descriptions is stored in a face image database. Each description is specified in face images, part data and keywords. A collection of stored descriptions is used for constructing latent semantic spaces, decision trees and association rules.

The system constructs a numeric latent semantic space and a combined latent semantic space. A numeric latent semantic space is constructed from only part data, and a combined latent semantic space part data and keywords, respectively. A procedure for seeking keywords consists of the following steps. At first, some similar face images to a given face image are sought using the numeric latent semantic space. Next, a centroid of the obtained face images is computed in the combined latent semantic space, which is used as a query. Keywords similar to

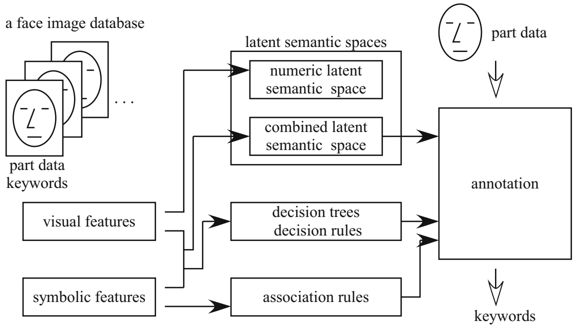


**Fig. 1.** An overview of annotation mechanisms and components of FIARS

this query are retrieved using the combined latent semantic space. Then, a cosine measurement is used. Retrieved keywords are seemed as keywords for the given face image.

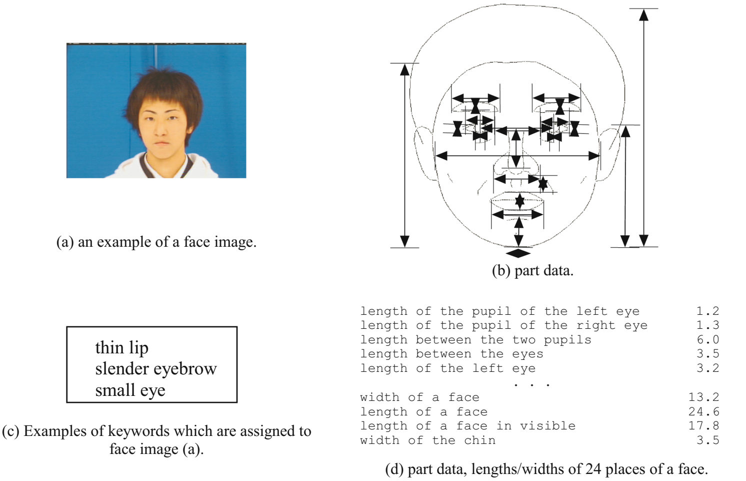
Decision rules represent whether a keyword is able to be assigned to a given face image, or not. On the other hand, association rules specify associations among keywords. When association rules are applied to a collection of keywords obtained by mechanisms of latent semantic spaces and decision trees, new keywords are captured by extending these keywords.

## 3 Face Description

An example of a face description is shown in Fig. 2. (a) shows an example of a face image. (b) presents part data. 24 places of the face parts are measured. (c) is a set of keywords which are assigned to (a). Keywords are restricted so that they represent sizes/shapes of face parts. Measurement places are set for measuring them. Moreover, when similar face images are retrieved using the numeric latent semantic space, another type of a feature, called point data, is considered [6]. The point data are given by distances between two points on the outline of a face. The number of measured places in the point data is greater than one in the part data. But, retrieval efficiencies of them are almost similar when each visual feature is used. Therefore, the part data are useful for their simplicity, and this feature is suitable to build decision trees.

Part data of face image $I_d$ are represented in a vector, $\boldsymbol{v}_d = (v_{d,1}, \ldots, v_{d,24})^T$, called a part vector. From this vector, a normalized part vector is obtained, $\boldsymbol{v}'_d = (v'_{d,1}, \ldots, v'_{d,24})^T$. $v'_{d,j}$ is normalized value of $v_{d,j}$, and given by $v'_{d,j} = (v_{d,j} - \mu_j)/\sigma_j + 1/2, (j = 1, 24)$. $\mu_j$ and $\sigma_j$ are the mean value and the standard derivation of face parts $j$, respectively.

On the other hand, 43 keywords are treated in current. For image $I_d$, keywords are represented in a keyword vector, $\boldsymbol{w}_d = (w_{d,1}, \ldots, w_{d,43})^T$. Each element $w_{d,j}$ is 1 or 0. They represent whether keyword $j$ is defined in face image $I_d$, or not, respectively.



(a) an example of a face image.

(b) part data.

thin lip
slender eyebrow
small eye

(c) Examples of keywords which are assigned to face image (a).

| | |
|---|---|
| length of the pupil of the left eye | 1.2 |
| length of the pupil of the right eye | 1.3 |
| length between the two pupils | 6.0 |
| length between the eyes | 3.5 |
| length of the left eye | 3.2 |
| . . . | |
| width of a face | 13.2 |
| length of a face | 24.6 |
| length of a face in visible | 17.8 |
| width of the chin | 3.5 |

(d) part data, lengths/widths of 24 places of a face.

**Fig. 2.** An example of a face description

# 4 Annotation Mechanisms

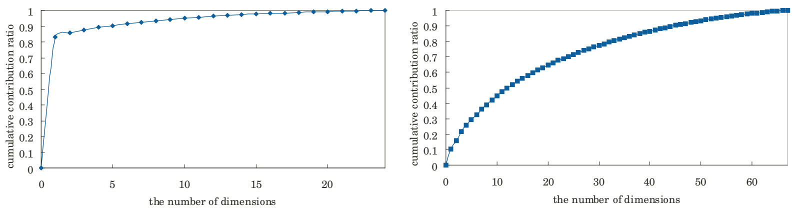
## 4.1 Dimensions of Latent Semantic Spaces

In latent semantic indexing, a matrix $A(m \times n)$ is decomposed into three matrices by the singular value decomposition [11], $A = U\Sigma V^T \cong U_k \Sigma_k V_k^T$. If the rank of $A$ is $r$, then $U$ is $m \times r$, the singular matrix $\Sigma$ is $r \times r$, and $V^T$ is $r \times n$. Let singular values be $\sigma_1, \ldots, \sigma_r$, and $\sigma_1 \geq \ldots \geq \sigma_r$. By selecting $k(1 \leq k \leq r)$, $A$ is approximated to $U_k \Sigma_k V_k^T$. On the other hand, a cumulative contribution ratio is computed as $\Sigma_{j=1}^{k} \sigma_j / \Sigma_{j=1}^{r} \sigma_j$. $k$ corresponds to the dimensions of a space. A contribution ratio seems to be useful because estimation of suitable dimensions is too hard [7].

To construct a numeric and a combined latent semantic spaces, matrix $\boldsymbol{N}$ and $\boldsymbol{C}$ are utilized, respectively. $\boldsymbol{N}$ is a collection of part vectors, $\boldsymbol{N} = (\boldsymbol{v}_1, \cdots, \boldsymbol{v}_d, \cdots, \boldsymbol{v}_n)$, where $n$ is the number of stored face descriptions. $\boldsymbol{C}$ is a collection of the face description in terms of keywords and part data, $\boldsymbol{C} = (\boldsymbol{c}_1, \cdots, \boldsymbol{c}_d, \cdots, \boldsymbol{c}_n)$. $\boldsymbol{c}_d$ is a concatenated vector $(\boldsymbol{w}_d; \boldsymbol{v}'_d)$, where $\boldsymbol{w}_d$ and $\boldsymbol{v}'_d$ are the keyword vector and the normalized part vector of $I_d$.

Figure 3 (a) and (b) show cumulative contribution ratios for the numeric latent semantic space and the combined latent semantic space, respectively. As shown in (a), it is seemed that individual part data are closely related each other. $\boldsymbol{N}$ can be reconstructed by a small number of dimensions in a sense that difference between the original matrix and the reconstructed matrix is little. As shown in (b), a large number of singular values are required for reconstructing $\boldsymbol{C}$ with little difference.

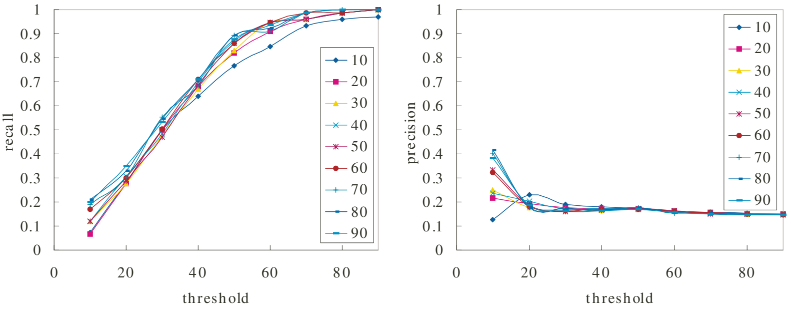
The dimension of each space is fixed in the developed system [6], which is equal to 3. Using these spaces we try to annotate new 10 face images. Figure 4 shows



(a) the cumulative contribution ratio in the numeric latent semantic space.

(b) the cumulative contribution ratio in the combined latent semantic space.

**Fig. 3.** Cumulative contribution ratios with respect to the numeric latent semantic space and the combined latent semantic space



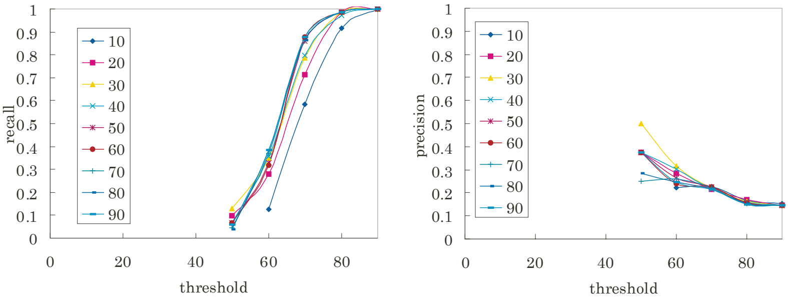
(a) recall.

(b) precision.

**Fig. 4.** Recall and precision when both dimensions of the numeric latent semantic space and the combined latent semantic space are equal to 3

recall and precision of retrieved keywords when thresholds are varied. (a) depicts recall, when a threshold for seeking keywords in the combined latent semantic space is changed from 10 to 90 degrees. The vertical axis and the horizontal axis are recall and a threshold for seeking keywords. At the same time, a threshold is changed for seeking similar face images in the numeric latent semantic space. Each line shows recall, when the thresholds for seeking similar face images are changed from 10 to 90 degrees. Moreover, (b) shows precision. When thresholds for keyword retrieval are increased, the recall are improved. The precision are low, although the thresholds are varied. When the thresholds for keyword retrieval are in during 30 and 90 degrees, precision are almost same although recall are improved. It seems that correct keywords are retrieved, but incorrect keywords are also retrieved as same as the correct ones.

Figure 5 shows recall and precision of retrieved keywords, when the dimensions of the numeric and the combined latent semantic space are equal to 3 and 33, respectively. The cumulative contribution of the combined latent semantic space is over 0.8 when the dimension is equal to 33. Thresholds for seeking faces and keywords are changed as described above. Under the thresholds less than 40 degrees for seeking keywords, keywords are not retrieved, since either similar face image or keyword is not retrieved. As shown in Fig. 4 and Fig. 5, the retrieval



(a) recall. (b) precision.

**Fig. 5.** Recall and precision when the dimension of the numeric latent semantic space is equal to 3 and the dimension of the combined latent semantic space is equal to 33

results shown in Fig. 5 are better than ones shown in Fig. 4. The recall shown in Fig. 5 are rapid increased, when the thresholds are increased. The precision shown in Fig. 5 are better than ones in Fig. 4, when the thresholds are around 50 degrees. To improve precision, it seems that the dimension of the combined latent semantic space is set high, and the threshold for seeking keywords is set low.

### 4.2 Decision Trees and Association Rules

Decision trees specify conditions to assign keywords in terms of part data. One decision tree consists of one root node, internal nodes, leaf nodes, and branches. The root node and the internal nodes are corresponding to individual part data, e.g., length of the pupil of the left eye, length between the two pupils, etc. Before a decision tree is constructed, simple discretization is applied to normalized part data. Their values are divided into three classes. About quarter of the stored face descriptions are assigned value ‘a’, half of them are assigned ‘b’, and the rest are assigned ‘c’ by ascending order. The values ‘a’, ‘b’ and ‘c’ are interpreted as ‘small/short’, ‘middle’ and ‘large/long’ depending on features.

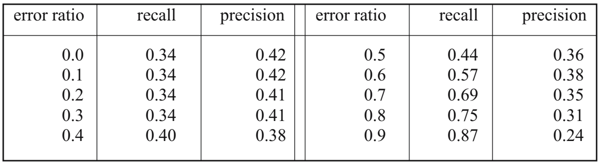
During construction of a decision tree, a node is tried to be expanded using entropy, i.e., information gain [5]. If all face descriptions indicated by the leaf node are positive examples in a sense that a certain keyword is assigned to all of them, the leaf node is not expanded. Moreover, pruning is applied using an error ratio. An error ratio is computed as *(the number of negative examples at a leaf node) / (a total number of face descriptions at a leaf node)*. The negative examples are the face descriptions that the keyword is not assigned to. If an error ratio of a leaf node is less than a specified error ratio, the node is not expanded.

A decision rule is built from a decision tree directly, represented in $A \rightarrow B$. $A$ is a set of patterns which are places of face parts. $B$ is a keyword. If part data of a given face satisfy $A$ then the keyword indicated in $B$ is assigned to the face. For example, when an error ratio is 0 the following rule is captured;

```
height_of_the_left_eyebrow(b) and width_of_the_face(b) and
height_of_the_right_eyebrow(a) and
distance_between_the_jaw_and_the_line_of_centers_on_eyes(c)
-> slender_eyebrow
```

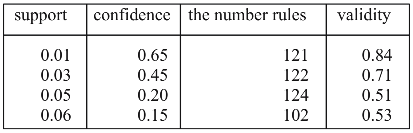
Table 1 shows recall and precision of captured keywords using decision rules, when an error ratio is changed. When an error ratio is small, the recall and the precision are almost constant. Although the recall is improved by increasing the error ratio, it is difficult to improve the precision.

**Table 1.** Recall and precision when an error ratio is changed



| error ratio | recall | precision | error ratio | recall | precision |
|---|---|---|---|---|---|
| 0.0 | 0.34 | 0.42 | 0.5 | 0.44 | 0.36 |
| 0.1 | 0.34 | 0.42 | 0.6 | 0.57 | 0.38 |
| 0.2 | 0.34 | 0.41 | 0.7 | 0.69 | 0.35 |
| 0.3 | 0.34 | 0.41 | 0.8 | 0.75 | 0.31 |
| 0.4 | 0.40 | 0.38 | 0.9 | 0.87 | 0.24 |

**Table 2.** Validity of captured association rules



| support | confidence | the number rules | validity |
|---|---|---|---|
| 0.01 | 0.65 | 121 | 0.84 |
| 0.03 | 0.45 | 122 | 0.71 |
| 0.05 | 0.20 | 124 | 0.51 |
| 0.06 | 0.15 | 102 | 0.53 |

On the other hand, an association rule $X \rightarrow Y$ represents co-occurrence relationships among keywords. $X$ and $Y$ are disjoint sets of keywords. To measure an association rule support and confidence are used [5]. For example, the following association rule is obtained;

```
small_eye and large_nose and dropping_eyes -> thin_lip
```

The support and the confidence of this rule are 0.04 and 0.79, respectively.

If the support is more than 0.01 and the confidence is more than 0.65, about 120 rules are obtained. Table 2 shows their validity. The validity of a rule is computed as *(the number of face descriptions including $X$ and $Y$ + the number of face descriptions including $X$ and which are suitable to assign $Y$) / (the number of face descriptions including $X$)*. If confidence is high, validity is also high. Therefore, it is considered that suitable keywords are able be obtained by checking the confidences of rules.

## 5 Concluding Remarks

Three mechanisms for face image annotation are presented; they are latent semantic indexing, decision trees and association rules. When these techniques are applied to the same face image, it occurs that inconsistent keywords are obtained. For example, two keywords which have opposite meanings are obtained.

In current, these techniques are used independently, and these subsystems have been developing individually. An annotator uses these subsystems, and annotation results have to be integrated carefully.

We plan to integrate these mechanisms, as future works. Moreover, relationships among keywords are defined, such as synonyms, antonyms, broader terms and narrower terms, like a thesaurus. By specifying such relationships, assignment of inconsistent keywords will be prevented. Furthermore, to semi-automatically determine the values of some parameters for working subsystems are required, e.g., dimensions of spaces, an error ratio, etc.

## Acknowledgement

We would like to thank Professor Iida, S. for helpful comments on this paper. The face images are used by permission of Softopia Japan, Research and Development Division, HOIP Laboratory. This research was supported in part by the Hibi Research Grant.

## References

1. Chellappa, R., Wilson, C.L., Sirohey, S.: Human and Machine Recognition of Faces: A Survey. Proceedings of the IEEE 83(5) (1995)
2. Datta, R., Ge, W., Li, J., Wang, Z.: Toward Bridging the Annotation-Retrieval Gap in Image Search. IEEE Multimedia (July-September, 2007)
3. Djeraba, C.: Association and Content-Based Retrieval. IEEE Tran. Knowledge and Data Engineering 15(1) (2003)
4. Fasel, B., Luettin, J.: Automatic Facial Expression Analysis: A Survey. Pattern Recognition 36(3) (2003)
5. Han, J., Kamber, M.: Data Mining, Concepts and Techniques. Morgan Kaufmann, San Francisco (2006)
6. Ito, H., Koshimizu, H.: Some Experiments of Face Annotation Based on Latent Semantic Indexing in FIARS. In: Gabrys, B., Howlett, R.J., Jain, L.C. (eds.) KES 2006. LNCS (LNAI), vol. 4252, pp. 1208–1215. Springer, Heidelberg (2006)
7. Kontostathis, A., Pottenger, W.M.: A Framework for Understanding Latent Semantic Indexing (LSI) Performance. Inform. Processing & Management 42 (2006)
8. Li, J., Wang, J.Z.: Real-Time Computerized Annotation of Pictures. IEEE Tran. PAMI (to appear, 2007)
9. Monay, F., Gatica-Perez, D.: Modeling Semantic Aspects for Cross-Media Image Indexing. IEEE Tran. PAMI 29(10) (2007)
10. Pantic, M., Rothkrantz, L.J.M.: Facial Action Recognition for Facial Expression Analysis From Static Face Images. IEEE Tran. SMC - Part B 34(3) (2004)
11. Skillicorn, D.: Understanding Complex Datasets. Data Mining with Matrix Decompositions. Chapman & Hall/CRC, Boca Raton (2007)
12. Softopia Japan Foundation: Face Image database, `http://www.hoip.jp/web=catalog/top.html`
13. Zhao, R., Grosky, W.I.: Narrowing the Semantic Gap? Improved Text-Based Web Document Retrieval Using Visual Features. IEEE Trans. on Multimedia 4(2) (2002)

# A Spam Filtering Method Learning from Web Browsing Behavior

Taiki Takashita, Tsuyoshi Itokawa, Teruaki Kitasuka, and Masayoshi Aritsugi

Department of Computer Science and Communication Engineering, Graduate School of Science and Technology, Kumamoto University, Kumamoto 860-8555, Japan
reo@st.cs.kumamoto-u.ac.jp,
{itokawa,kitasuka,aritsugi}@cs.kumamoto-u.ac.jp

**Abstract.** In this paper a spam filtering method is proposed. We focus on user behavior that most email users browse the Web. The method reduces troublesome maintenance of the spam filter, since the filter learns from Web browsing behavior in the background. The method uses Web browsing behavior of each user to learn ham words. Ham words are picked up from browsed Web pages using TF-IDF and stored in the database called ham words list. For each received email, the method extracts keywords from the email, including Web pages of the URLs. If some keywords are in the ham words list, the email is treated as a ham. In our experiments, several spam emails which cannot be detected by a Bayesian filter are detected as spams.

## 1 Introduction

Spam email accounts for 90 to 95 percent of all email in 2007, up from an estimated five percent of email in 2001, and spam emails become the worse form of junk advertising than postal junk mails and telemarketing calls[1]. Spam filtering is required for not only technical reasons such as overspend the network bandwidth and email storage, but also social issues such as child safety, phishing email, and so on. We are already hard to find ham emails without a kind of anti-spam technologies.

The major anti-spam technologies are categorized into sender-side technologies or receiver-side technologies. The filtering methods which this paper concerns are categorized the latter. The former is to prevent spammer from sending email. Outbound port 25 blocking of ISP is an example of the sender-side technologies.

In this paper, we focus on user behavior that most email users also browse the Web. Conventional spam filters use information extracted from emails. The proposed method learns the user preference from Web browsing behavior. The merit of the method is reduction of maintenance task of the filter, since it learns the user preference in background of browsing behavior. We show the basic concept and design of our method, and results of preliminary experiments in this paper.

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 774–781, 2008.

This paper is organized as follows. In Section 2, related work is introduced. The proposed method is described in Section 3. In Section 4, evaluation results of the proposed method are shown. Finally we conclude the paper in Section 5.

## 2 Related Work

Anti-spam research and development is an ongoing battle with both spammers and spam fighters. It's becoming ever more sophisticated[2]. There are many researches of spam filtering. Work on spam filtering can be divided into two categories: content-based approach and collaborative approach. On the content-based approach, the classification of an email is based on an analysis of the content of the email. The second is collaborative approach, which depends on the collaboration of groups of users to share information about spam[3].

Both approaches are widely used in these days. Collaborative filtering is employed by Web mail service providers. Bayesian filtering, which is a kind of content-based approach, is mostly built in MUA (mail user agent).

The collaborative filtering is a server-based approach. It shares spam information between many users. Once a user reports a received email as a spam to the server, the server updates information of spam. Then, other users will see the email that has the same content with a spam flag or in a spam folder. This filtering is in broad category of folksonomy. Since the filter is maintained globally, unique preference of each user is hard to be reflected into the filter. To make custom filter for each user, the filter maintenance is needed somewhat by the user as same as Bayesian filter.

Bayesian filter is a content-based approach[4]. Many MUAs adopt the filter to detect spam. Basic concept of Bayesian filter is based on Bayesian combination of the spam probabilities of individual words in an email. All received emails are classified into spam or ham, according to the threshold of the probabilities. To classify emails, the filter has corpuses of spams and hams. These corpuses are maintained by users themselves, since the probability of false classification depends on them.

In addition, there are many kinds of content-based approaches. In [5], the method that processes email messages as image data is proposed. When we manually filter the spam, we glance at a message as an image instead of reading it carefully. The method detects spam by transforming a received email into image in accordance to HTML tag structure.

Note that the maintenance of these filtering methods is very tedious and expensive task, since it usually takes long time to get the necessary information for the maintenance. In this paper, we propose a novel method that gets information from Web browsing.

## 3 Our Spam Filtering Method

### 3.1 Structures

On an assumption that most of the email users browse Web pages, we try to examine the spam filtering method which learns the user's preference from the Web browsing behavior. The user's preference of the email will be quite similar to that of the browsed Web pages. The proposed method provides an individual filter to each user. The filter has a database called ham words list, which is created from the Web browsing behavior of the user. The preference varies according to an interest of each user. Therefore we think that the content of the browsed Web pages is suitable to learn the preference of the user. Our approach was inspired by [10], in which an information management assistant gathers contextual information from user interactions and leverages it to support just-in-time information access.

The method needs to develop the interface between spam filter and the browser. However the method makes many users free from some part of the maintenance of the Bayesian filter or collaborative filtering systems.

The method consists of two stages: the first stage is a creation of ham words list, and the second stage is a filtering of received emails with ham words list. Fig. 1 shows the outline of the stages. At the first stage, when the user browses Web pages day-by-day, the HTML sources of the pages are gathered to extract the ham words. The words in the HTML source are processed according to TF-IDF (Term Frequency–Inverse Document Frequency)[6]. For each page, some words which have high TF-IDF score are stored in the ham words list.

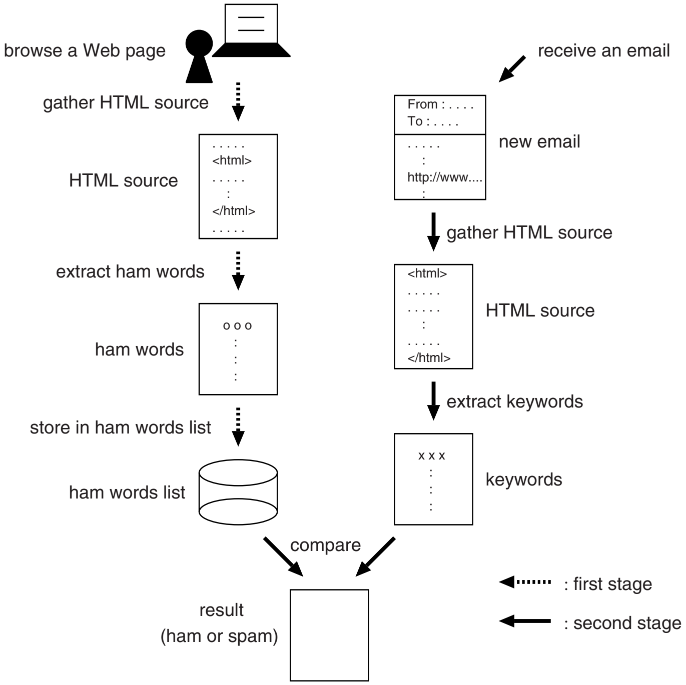


**Fig. 1.** Overview of the proposed method

At the second stage in Fig. 1, the method filters the received email with the ham words list created in the first stage. To judge the email, we have to pick up the words from the received email. The filter refers the one or more URLs in the email to retrieve keywords of the email. The flow of retrieval of keywords is similar to the first stage. Finally to determine the email is a ham or a spam, relevancy of the email to the user preference is tested through matching of ham words list and keywords of the email. The detail of each stage is described in the remaining part of this section.

### 3.2 The Ham Words List of Web Browsing Behavior

At the first stage, the ham words list is updated according to the Web browsing behavior of each user. The ham words list is used in the second stage to determine the received email is a ham or a spam.

The ham words list will be updated using Web pages that the user browses. For each Web page, HTML source of the page is processed to find new ham words. Words in the HTML source are weighted by TF-IDF. TF-IDF score of a word $t_i$ in the HTML source $d_j$ is defined as follows:

$$\begin{aligned}\text{TF-IDF}_{i,j} &= \text{TF}_{i,j} \times \text{IDF}_i \\ &= \frac{n_{i,j}}{\sum_k n_{k,j}} \times log \frac{|D|}{|\{d : d \ni t_i\}|}\end{aligned}$$

where $n_{i,j}$ is the number of occurrence of the word $t_i$ in document $d_j$, $|D|$ is the total number of documents, and $|\{d : d \ni t_i\}|$ is the number of documents in $D$ which contain the word $t_i$.

Currently TF-IDF$_{i,j}$ of each word $t_i$ is calculated using Yahoo! API[7]. $|D|$ is treated as the total number of sites which Yahoo crawls. $|\{d : d \ni t_i\}|$ is treated as the number of sites returned from Yahoo! contextual Web search for the word $t_i$.

$n_{i,j}$ of TF-IDF can be calculated only from the HTML source of the Web page. In Section 4, we use emails and Web pages written in Japanese. Before we calculate TF-IDF score of each word, morphological analysis for Japanese language is required. We employ a tool of Japanese language morphological analysis called Sen[8]. Only common nouns and proper nouns are the candidates of words to add the ham words list.

Through preliminary evaluation, we find exceptional words which should not be treated as ham words in the list. When one of exceptional words is included in a ham words list, many false negatives occur. False negative means that a spam email is not detected as a spam. Exceptional words is selected heuristically. In the experiments described in Section 4, we used 20 words as exceptional words, which are Japanese nouns of “search”, “register”, “free”, “member”, “site”, “image”, “login”, “password”, “point”, “year”, “category”, etc.

### 3.3 Filtering Received Emails

When a user or a mail server receives an email, the email is judged by the filter which uses the ham words list. The filter can be built into either SMTP server or MUA.

At first, the filter extracts the keywords of the email. In this work, we assume that keywords are extracted from the Web pages linked by URLs in the email body. The keywords of the email are selected according to the same policy of selection of ham words. The top $k$ words of TF-IDF score are selected as keywords of the email. $k$ is varied from 2 to 6 in Section 4.

To judge an email, the filter calculates conformance of the keywords of the email with the words in the ham words list. In the preliminary evaluation described in Section 4, we employ a simple calculation, i.e., if the keywords of the email are contained in the ham words list more than or equal to a threshold number, the email is judged as a ham. Otherwise, it is judged as a spam. The number of keywords is varied in Section 4. We will consider more sophisticated conformance calculation in future.

## 4 Experiments

### 4.1 Preliminary Evaluation

The proposed method is evaluated through the following environments of a virtual user. We assume that the user has interests in eight categories: childcare, corporate stock, horse races, movies, fortune-telling, news of show business, recipes, and Internet auction. The ham words list was created by 838 Web pages, including about 300 pages of the above categories. 1,000 emails were used as target emails: 500 emails were hams, and the other 500 emails were spams. All emails are picked up from email magazines and actual spams received by the authors. Ham emails are from the magazines of the categories of user's interest. Spam emails are actual spams and emails categorized into giveaway items, point programs, and adult of email magazines.

Three cases of experimental results are shown in Table 1. Both the number of keywords picked up from each email and threshold of judgement of ham are increased and decreased simultaneously. In all cases, the threshold is a half of the number of keywords. For example of case 1 in Table 1, two keywords are picked up from each email. If at least one of the keywords is found in the ham words list, the email is judged as ham. From Table 1, no ham is judged as a spam (false positive), and 270 spams are judged as hams in this case. In the comparison of keywords and the list, the TF-IDF score of each keyword is not referred in this evaluation.

In this experiment, unfortunately, our method did not achieve good results, since the number of false negatives was very high. The following is the short discussion of the results. The important requirement of spam filtering is low probability of false positive occurrence. False positive is the misjudgement of a ham as a spam. The method can probably stand for the requirement, since

**Table 1.** Number of errors using the proposed method

| Case | Parameter | | Results | |
|---|---|---|---|---|
| | # of keywords | Threshold | # of false positive | # of false negative |
| 1 | 2 | 1 | 0 (0%) | 270 (54%) |
| 2 | 4 | 2 | 8 (1.6%) | 251 (50.2%) |
| 3 | 6 | 3 | 7 (1.4%) | 240 (48%) |

**Table 2.** Number of errors using Bayesian filter

| | # of false positive | # of false negative |
|---|---|---|
| Bayesian filter (Thunderbird) | 195 (39%) | 25 (5%) |
| Conjunction | – | 2 (0.4%) |

no false positive occurred in the case of low threshold. False positive implies a kind of lost of emails, when the case of that the filter brings the ham into a quarantine (i.e. spam folder) instead of inbox. Nowadays user may receive over 100 spam per day. It's hard to find a few ham in a quarantine which contains a large number of spams.

On the other hand, false negative occurred in high probability around 50%. The results show that the method is not enough to judge spam precisely. However the disadvantage will be avoided by combination of conventional spam filters. In Section 4.2, we will discuss this combination.

### 4.2 Comparison with a Bayesian Filter

The method is compared with the spam filter based on Bayesian filter. As a Bayesian filter, we used Thunderbird[9] that is the popular MUA with Bayesian spam filter. In the experiment, Bayesian filter learns 1,000 ham and 1,000 spam emails before filtering. These emails were selected from archive of email magazines of the same categories in Section 4.1. 1,000 target emails are used as same as Section 4.1.

Table 2 shows the number of false positives and false negatives using Thunderbird. Bayesian filter of Thunderbird misclassifies 39% of ham emails into spams, and 5% of spams into hams.

Firstly, we discuss the false negative which is misjudgement of a ham. False negative using Bayesian filter occurs for 25 emails. These emails contain very short text in mail body and URL. Fig. 2 (a) and (b) are typical examples of the email of false negative by Bayesian filter (the message is written in Japanese). It contains only 5 sentences, two URLs, and telephone number in the body. This kind of email is hard to detect as a spam by current Bayesian filter.

Applying the proposed method to these 25 false negative emails, 23 emails can be judged as spams. For all 25 emails, words in the email body is hard to judge correctly. However the method can pick up keywords of the email from the HTML source of URLs in the email body. The results show that the proposed

```
おはようございます。
今井さやかちゃん出勤です♪
http://m.garden-gal.com/index.php?ac=gal_detail&cp=652
午前中のご予約に空きがでました !!
お急ぎおといあわせください。
・ 0354285131

解除するには下記 URL にアクセスして下さい。
http://www.emaga.com/tool/automail.cgi?code=garden01&mail=(omitted)
```

(a) mail body

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN">
<html><head><title>渋谷 ホテル型性感 クラブ ガーデン/在籍ギャルデータ</title>
<meta name="keywords" content="渋谷, 渋谷駅, 道玄坂, 百軒店, ヘルス, 派遣型ファッションヘル
ス, ホテルヘルス, ホテヘル, 性感マッサージ,AV 女優,OL, 学生, ギャル, マット,AF">
<meta name="description" content="ClubGarden は渋谷、渋谷駅発のホテルヘルス、ファッションヘル
スです。現役 AV 女優、OL、学生、女子大生、在籍ギャル多数!!詳細な在籍ギャル達のプロフィール等を掲載し
ています。当店では、性感コース、AF コースをご用意しています。"></head><body bgcolor="#daebfc">
<div align="center">
アイドル級のかわいさです !! <br>
<img src="TempImgs/68b3b12e8eebdf96db847b066fea7254.jpg" border="0"><br>S
<a href="/index.php?sessionid=12040913369261951&amp;ac=gal(omitted)</a><br></div>
<hr size="1">
<font color="#bffda9">●</font>今井さやか<font color="#bffda9">●</font><br>
19 歳 T.160<br>
B88(F)W57H85<hr size="1">
ルックス最高レベルです、恋しちゃいます !! <hr size="1">
<a href="/index.php?sessionid=12040913369261951&amp;ac=gal_profile&amp;cp=652">詳
細</a><br>
<a href="/index.php?sessionid=12040913369261951&amp;ac=gal_shift&amp;cp=652">出勤予定
表</a>
<hr size="1">
  営業時間<br>
9：00～23：50<br>
<a href="tel:0354285131">03-5428-5131</a>
<hr size="1">
<a href="/index.php?sessionid=12040913369261951&amp;ac=top">へ</a><br>←戻る
<hr size="1">
<font size="1">
<center>&#169;2007-2008 ClubGarden. All Rights Reserved.</center>
</font>
</body></html>
```

(b) HTML source of first URL in mail body

**Fig. 2.** An example email of false negative (written in Japanese)

method can cover a weakness of Bayesian filter. To reduce false negative, we can reexamine the email by the proposed method, after an email is determined as a ham by Bayesian filter.

Secondly, we discuss the false positive using only Bayesian filter. The results will not be used for explaining ineffectiveness of Bayesian filter. High probability of false positive of Thunderbird is probably caused by the learning environment of this experiment. All learned hams are from email magazines. For an email magazine, all emails of this magazine are completely judged as spams. There are the following two types of false positive emails.

– many verbose lines of advertisement in the mail body.
– high variability of keywords, e.g., in the category of news of show business there are very wide variety of keywords.

We pick up some false positive emails to analyze the proposed method. The keywords of each email selected by the proposed method are right keywords in each category.

By comparing with Bayesian filter, we conclude that the proposed method has effective situations which are hard to adapt the Bayesian filter.

## 5 Conclusion

We proposed a spam filtering method that uses Web browsing behavior in this paper. The method retrieves the preference of each user through Web browsed Web pages. We reported preliminary results of experiments. The results show that several spams which Bayesian filter cannot classify as spams can be judged as spams. These spams seem to be hard to classify precisely by Bayesian filter, since they contain a short body of email such as a few sentence and URLs.

As future work, we will consider the scheme to combine with other filters such as Bayesian filter. To combine several filters, we have to manage discrepancy between judgements of filters. The experiments with sophisticated data set such as [11] are also included in our future work.

## References

1. Barracuda Networks, Inc.: Barracuda Networks Releases Annual Spam Report (press release, 2007), http://www.barracudanetworks.com/ns/news_and_events/index.php?nid=232
2. Goodman, J., Cormack, G.V., Heckerman, D.: Spam and the Ongoing Battle for the Inbox. Communication of ACM 50(2), 24–33 (2007)
3. Cunningham, P., Nowlan, N., Delany, S.J., Haahr, M.: A Case-Based Approach to Spam Filtering that Can Track Concept Drift. In: Proc. ICCBR 2003 Workshop on Long-Lived CBR Systems (2003)
4. Graham, P.: A Plan for Spam (2002), http://www.paulgraham.com/spam.html
5. Kumagai, N., Aritsugi, M.: On Applying an Image Processing Technique to Detecting Spams. In: Proc. 21st International Conference on Data Engineering Workshops (ICDEW 2005), p. 1172 (2005)
6. Salton, G., Buckley, C.: Term-Weighting Approaches in Automatic Text Retrieval. Information Processing and Management 24(5), 513–523 (1988)
7. Yahoo! Inc.: Yahoo! Search Web Services, http://developer.yahoo.com/search/
8. http://ultimania.org/sen/ Sen (in Japanese)
9. Mozilla: Thunderbird, http://www.mozilla.com/thunderbird/
10. Budzik, J., Hammond, K.J.: User Interactions with Everyday Applications as Context for Just-in-time Information Access. In: Proc. 5th International Conference on Intelligent User Interfaces, pp. 44–51 (2000)
11. Androutsopoulos, I., Koutsias, J., Chandrinos, K.V., Spyropoulos, C.D.: An Experimental Comparison of Naive Bayesian and Keyword-based Anti-Spam Filtering with Personal E-mail Messages. In: Proc. 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2000), pp. 160–167 (2000)

# A Method of Graphics Composition Using Differential SVG Documents

Motohiro Matsuda[1], Kazunari Ito[2,4], Martin J. Dürst[3], and Kôiti Hasida[4]

[1] Graduate School of Science and Engineering, Aoyama Gakuin University
[2] School of Social Informatics, Aoyama Gakuin University
[3] College of Science and Engineering, Aoyama Gakuin University
[4] National Institute of Advanced Industrial Science and Technology

**Abstract.** This paper describes a novel method extracting editing differences between pairs of SVG graphics documents. These differences are extracted based on the analysis of the tree structure of SVG, and are generalized in order to abstract from the specifics of targets and document structure. The generalized differences can then be applied to other SVG graphics, resulting in new, heretofore unavailable graphics. We show the effectiveness of our method with experiments involving a variety of SVG documents.

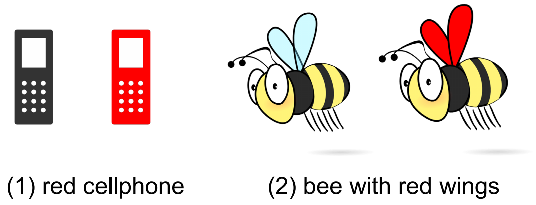
## 1 Introduction

XML is used widely for document and data formats. SVG (Scalable Vector Graphics) [1] is a 2D vector graphics format based on XML, gaining in popularity recently. SVG expresses a graphic by a combination of primitive graphic elements such as line, rectangle, ellipse, and polygon, and uses styling attributes to indicate visual properties such as color or transparency. We use the XML structure of an SVG document to extract editing differences between pairs of graphics.

The motivation for our research is coming from our work on pictograms. Pictograms are strongly simplified graphics used to express objects and ideas. Pictograms are increasingly used for intercultural communication, learning assistance for infants and livelihood assistance for persons with disabilities. Our research has focused on expressing sentence-level concepts by arranging several pictograms in 2D space [2], and to study the relationship between 2D arrangements and semantic relationships [3]. To support the creation of new pictograms and the collection of data, we also investigated pictogram authoring [4], allowing the combination of existing pictograms.

In many cases, however, the simple combination of existing pictograms is not sufficient to express a new idea; some changes to the pictogram components are necessary. On the other hand, once changed pictograms are available, it should be possible to abstract changes from the actual pictograms and to reapply them to other pictograms. Such operations should not be limited to pictograms, but should be applicable to more complex graphics such as illustrations, as long as the editing operations themselves do not exceed a certain level of complexity.

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 782–789, 2008.



(1) red cellphone (2) bee with red wings

**Fig. 1.** Reuse of "red" data

As an example, Fig. 1 on the left shows a (black) "cellphone" and a "red cellphone", the later the result of the user editing the cellphone's body to red. It should be possible from these two images to extract the data corresponding to "cellphone" and "red", and to reuse this data in different contexts. As an example, the data for "red" extracted from the two cellphone graphics should be applicable to the "bee"[1] on the right of Fig. 1, resulting in a "bee with red wings".

In this paper, we describe how to extract such data using editing differences between pairs of SVG graphics, and how to apply such differences to the generation of new SVG graphics from existing graphics.

## 2 Method

### 2.1 Outline

Simple editing operations on graphics can be analyzed based on the state of the graphic before and after editing. Also, such editing can be decomposed into an *editing target* and an *editing operation*. The editing target expresses the object to which the editing operation is applied. The editing operation expresses the difference between the non-edited document and the edited document. We call the combination of non-edited and edited graphics, the editing target and editing operation a *differential component*.

In this paper, editing operations may also include the addition or a change of an animation. Extracting animations from sequences of raster images is difficult process [5]. However, in SVG, animations are expressed as subelements of the object which is being animated. We can therefore handle animations in the same way as we handle other editing operations, as operations on an XML tree.

### 2.2 Input

The input data for our method consists of SVG graphics. In general, three kinds of data are necessary, (1) non-edited graphic, (2) editing target, and (3) edited graphic. In Fig. 2, the editing target identifies the "wings" objects in the non-edited graphic. The editing target does not need to be a single object. It can e.g. consist of two wings, or four legs. In many cases, the editing target can be

[1] from Open Clip Art Library, http://www.openclipart.org/

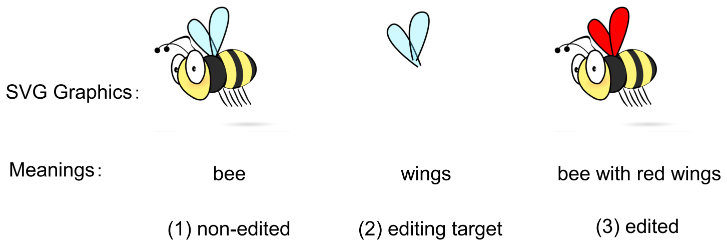


**Fig. 2.** An Example of Input

```
<xd:xmldiff>
  <xd:node match='5'>
    <xd:node match='1'>
      <xd:node match='10'>
        <xd:change match='@style'>fill:#ff0000;stroke:#000000;stroke-width:6;</xd:add>
      </xd:node>
      <xd:node match='12'>
        <xd:change match='@style'>fill:#ff0000;stroke:#000000;stroke-width:6;</xd:add>
      </xd:node>
    </xd:node>
  </xd:node>
</xd:xmldiff>
```

**Fig. 3.** An Example of Extracted Differential SVG Source (simplified)

extracted automatically from the non-edited and edited documents. However, this is not always the case. For adding operations, the target (the object being added to) and the object being added may not be in a direct relationship in the XML tree. To simplify processing, we therefore assume that the editing target is explicitly identified.

### 2.3 Differential Extraction from SVG Tree Structure

Editing differences are represented by differences in XML tree structure between the non-edited and edited SVG graphics. Detecting differences between two XML documents is a well-researched topic, see [6]. However, a standard format expressing these differences has not yet emerged. We use the Microsoft XML Diff and Patch Tool [7] to extract the differences between two SVG graphics, and the corresponding XML Diff [8] format to represent the result.

The result of extracting the difference between graphics (1) and (3) in Fig. 2 is shown in Fig. 3. The XML Diff format represents tree editing operations as insertions, changes, and deletions, using the `xd:add`, `xd:change`, and `xd:remove` elements, respectively. Movement of subtrees is represented by a pair of `xd:add` and `xd:remove` elements. The target of an editing operation is defined by the hierarchy of the `xd:node` elements and their `match` attributes. The two targets in Fig. 3 can be expressed by the following two XPath expressions: `node[5]()/node()[1]/node()[10]` and `node[5]()/node()[1]/node()[12]`. Please note that the actual numbers will differ because XPath counts text nodes, whereas XML Diff does not. These two editing target nodes represent the "bee's wings" in Fig. 2.

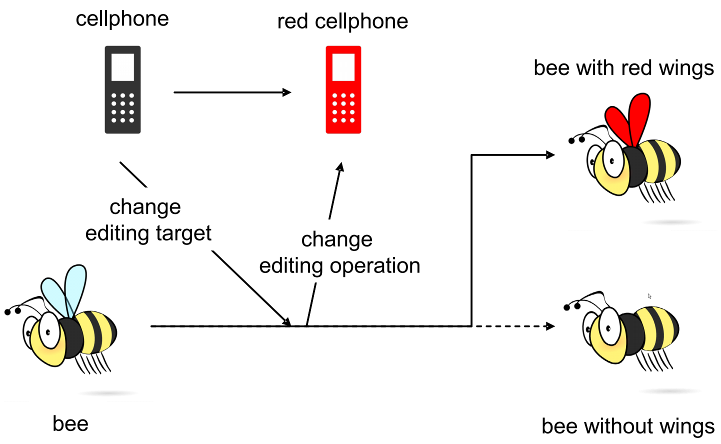


**Fig. 4.** An Example of Synthetic SVG Document

So, this difference means that it changes Fig. 2 (1) "bee" into (3) "bee with red wings" by applying editing operations to the two editing target nodes.

The reader should be aware of the fact that the target nodes identified here may not correspond one-to-one with the edited objects as seen by the user. It is possible that an object as seen by the user is expressed as a combination of graphics primitives in SVG, or that several objects, e.g. all the buttons on the cellphone in Fig. 1, are represented as a single group or path object in SVG.

Editing targets often correspond to graphic components such as "wings" or "antennae", while editing operations often correspond to presentation attributes such as colors. The separation of editing targets and editing operations allows creating a variety of new SVG documents by combining editing targets and editing operations from different differential components. For example, in Fig. 2, if the editing target "wings" is replaced by "antennae", we end up with a "bee with red antennae". In a similar way, if the editing operation "red" is replaced by "without", the result is a "bee without wings".

We can not only use different editing targets and editing operations from the same original graphic, but can also apply these across different graphics. Fig. 4 shows an example. The editing operation "red", extracted from the differential component between "cellphone" and "red cellphone", is combined with the editing target "bee's wings", extracted from the differential component between "bee" and "bee without wings", and applied to "bee" to result in a new SVG graphic, "bee with red wings".

## 3 Experimental Results and Discussion

### 3.1 Outline

In our experiment, we try to synthesize new graphics based on a pair of differential components. Overall, we use eight differential components. Figure 5

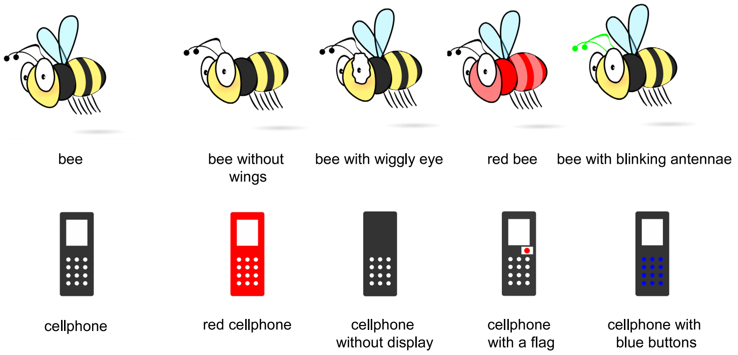


**Fig. 5.** SVG Documents used for the Experiment

**Table 1.** Problems used in Experiment

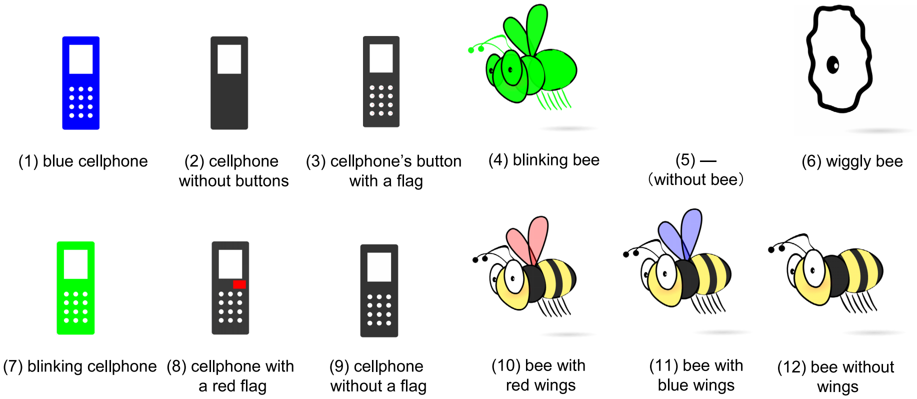
| number | source of original graphic and editing target | | source of editing operation | |
|---|---|---|---|---|
| | edited document | editing target | edited document | editing target |
| 1 | red cellphone | cellphone | cellphone with blue buttons | buttons |
| 2 | cellphone with blue buttons | buttons | cellphone without display | display |
| 3 | cellphone with blue buttons | buttons | cellphone with a flag | cellphone |
| 4 | red bee | bee | bee with blinking antennae | antennae |
| 5 | red bee | bee | bee without wings | wings |
| 6 | red bee | bee | bee with wiggly eye | eye |
| 7 | cellphone with a flag | cellphone | bee with blinking antennae | antennae |
| 8 | cellphone with a flag | flag | red bee | bee |
| 9 | cellphone without a flag | cellphone | bee without wings | wings |
| 10 | bee without wings | wings | red cellphone | cellphone |
| 11 | bee without wings | wings | cellphone with blue buttons | buttons |
| 12 | bee without wings | wings | cellphone without display | display |

shows two non-edited graphics on the left and eight edited graphics on the right, creating eight differential components.

Table 1 shows the twelve pairs of differential components used for our experiments. The differential component on the left provides the original graphic as well as the target object, while the differential component on the right provides the editing operation. For example, in Problem 1, we create a "blue cellphone" by using the editing target from the differential component between "cellphone" and "red cellphone" and the editing operation from the differential component between "cellphone" and "cellphone with blue buttons".

### 3.2 Synthesis Results

The synthesis results are shown in Fig. 6. Each graphic in Fig. 6 corresponds to a row in Table 1. In Problem 3, each button is decorated with a little flag, but these flags are hardly visible. Problems 4 and 7 use animation, which is not visible in print. The result of Problem 5 is an empty graphic, as expected.



(1) blue cellphone
(2) cellphone without buttons
(3) cellphone's button with a flag
(4) blinking bee
(5) — (without bee)
(6) wiggly bee
(7) blinking cellphone
(8) cellphone with a red flag
(9) cellphone without a flag
(10) bee with red wings
(11) bee with blue wings
(12) bee without wings

**Fig. 6.** Synthesis Result

### 3.3 Discussion

Overall, our results are meeting our expectations in most cases. The exception is Problem 6. This graphic is synthesized by applying the wigglyness of the bee's eye to the bee itself. Because in SVG, "wiggly" is not expressed as an attribute, but as part of the path data of the overall shape, the bee is replaced by an enlarged eye. This shows the current limitations of our method.

## 4 Related Research

As far as we know, our approach to extracting differences from structured graphic data such as SVG graphics is new. There is a lot of research on extraction of objects and knowledge from images, mainly in the context of CBR (Content Based Retrieval) [9]. CBR extracts the objects from their characteristics such as color or shape. QBIC [10] and VisualSEEk [11] are some representative examples of CBR systems. These systems adopt a non-heuristic method to extract objects from raster images.

For vector graphics, there are contour extraction methods using figurative elements [12], and feature-point matching [13]. These approaches extract meaningful objects from clearcut images such as cliparts or pictograms, but have difficulties with images with unclear contours. Kushima et al. [14] study the use of an image database to try to improve precision, but complete automation of object extraction from images is still difficult.

The reason for this difficulty is not only the lacking precision of the extraction methods, but also human subjectivity and sensitivity. We think that semiautomatic techniques involving human assistance can solve this problem effectively and efficiently [15] [16].

## 5 Conclusions and Future Work

In this paper, we focused our attention on SVG graphics edited by humans. Starting with non-edited and edited data, we showed how it was possible to extract various kinds of data and reuse it to create new graphics. We verified the effectiveness of our method with some experiments. The key idea of our method is to use editing differences rather than single graphics. Our method is easy to use because it does not require complicated preprocessing. Using our method in the context of our Pictogram Authoring system should help making the generation of pictograms easier.

Our method is still not satisfactory when concepts are expressed as changes to the shape itself rather than as separate attributes or elements. Solving this issue will require some more work. We plan to apply our method to analyze or improve the structure of SVG objects throughout the history of their editing process.

We have described our method in terms of data extraction. However, on some level, the data corresponding to "wing" or "red" is a representation of the corresponding concept. We plan to explore this idea of concept extraction using differential documents in more detail. In addition, we intend to examine rule extraction from a large number of differential SVG documents, because we think that general concepts should not be based on a single editing operation, but on wide input from actual users. This can significantly increase the chance to be able to identify meaningful units and concepts in a graphic.

## References

1. Scalable Vector Graphics (SVG), http://www.w3.org/Graphics/SVG/
2. Ito, K., Hasida, K.: Ontology mapping to promote making and understanding pictograms. DBSJ Letters (in Japanese) 5(2), 93–96 (2006)
3. Matsuda, M., Ito, K., Dürst, M.J., Hasida, K.: A Rule Extraction Method for Aarranging Pictogram Components. DBSJ Letters (in Japanese) 6(1), 165–168 (2007)
4. Ito, K., Matsuda, M., Dürst, M.J., Hasida, K.: SVG Pictograms with Natural Language Based and Semantic Information. In: Proceedings of the 5th International Conference on Scalable Vector Graphics (SVG Open 2007) (2007)
5. Anandan, P., Irani, M., Kumar, R., Bergen, J.: Video as an image data source: efficient representations and applications. In: Proceedings of the 1995 International Conference on Image Processing, p. 318 (1995)
6. Peters, L.: Change detection in xml trees: a survey. In: 3rd Twente Student Conference on IT (2005)
7. Microsoft XML Diff. and Patch 1.0, http://apps.gotdotnet.com/xmltools/xmldiff/
8. XML Diff., http://schemas.microsoft.com/xmltools/2002/xmldiff
9. Marsicoi, D., Cinque, L., Levialdi, S.: Indexing pictorial documents by techniques. Image and Vision Computing 15, 119–141 (1997)
10. Flickner, M., et al.: Query by image and video content: the QBIC system. Computer 28(9), 23–32 (1995)

11. Smith, J.R., Chang, S.-F.: VisualSEEk: A fully automated content-based image query system. ACM Multimedia, 87–98 (1996)
12. Hayashi, T., Onai, R., Abe, K.: Vector image segmentation for content-based vector image retrieval. In: CIT 2007: Proceedings of the 7th IEEE International Conference on Computer and Information Technology, pp. 695–700 (2007)
13. Kim, B., Yoon, J.P.: Similarity measurement for aggregation of spatial objects. In: SAC 2005: Proceedings of the 2005 ACM symposium on Applied computing, pp. 1213–1217. ACM, New York (2005)
14. Kushima, K., Akama, H., Kon'ya, S., Yamamuro, M.: Exsight: Highly accurate object based image retrieval system enhanced by redundant object extraction. In: Lu, H., Zhou, A. (eds.) WAIM 2000. LNCS, vol. 1846, pp. 331–343. Springer, Heidelberg (2000)
15. Ashley, J., Barber, R., Flickner, M., Hafner, J., Lee, D., Niblack, W., Petkovic, D.: Automatic and semiautomatic methods for image annotation and retrieval in query by image content (QBIC). In: Proc. Storage and Retrieval for Image and Video Database III, vol. 2420, pp. 24–35 (1995)
16. Mohan, A., Papageorgiou, C., Poggio, T.: Example-based object detection in images by components. IEEE Transactions on Pattern Analysis and Machine Intelligence 23(4), 349–361 (2001)

# Externalization Support of Key Phrase Channel in Presentation Preparation

Koichi Hanaue and Toyohide Watanabe

Department of Systems and Social Informatics,
Graduate School of Information Science, Nagoya University
Furo-cho, Chikusa-ku, Nagoya, 464-8603, Japan
{hanaue,watanabe}@watanabe.ss.is.nagoya-u.ac.jp

**Abstract.** In this paper, we propose a framework of supporting presentation scenario composition from a viewpoint of externalizing a presenter's intention. Traditional presentation slides in the style of bullet lists lack in rhetorical relations between idea fragments. This makes it difficult that audiences could understand presentation utterances and presentation slides. In order to solve this problem, we introduce a concept of logical frame. The logical frame is a semantic block in which idea fragments are organized rhetorically. A presentation scenario is composed through the process of organizing idea fragments within a logical frame and relating logical frames to each other. Based on this model of presentation scenario, we have implemented a scenario composition interface.

**Keywords:** presentation scenario, logical frame, knowledge externalization, script.

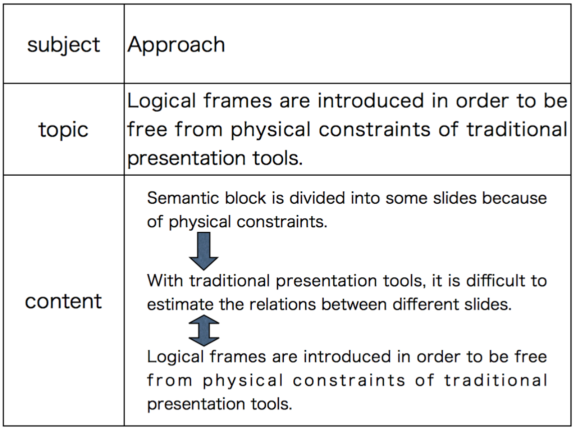
## 1 Introduction

Presentation is one of the most important means for transferring knowledge. In most presentations, presentation tools such as Apple Keynote[1] and Microsoft PowerPoint[2] have been used. These tools enable us to make presentation processes effectively with colorful graphics and animations.

However, the traditional styles of presentation preparation have some problems. First, presenters tend to spend more time and effort in preparing for presentation than in considering presentation scenarios. Second, most presenters use the standard formats of presentation slides set up by presentation tools, and thereby rely on the style of phrase headlines with bullet lists. This style makes the discussion point of each slide unclear[3]. For example, Tufte[4] points out the same thing on the assessment report of the accident in the space shuttle Columbia. He argued that the important information on a presentation slide is shadowed in bullet lists.

As Tufte argues, the traditional presentation tools lack in the functions of handling the intention of a presenter: especially, the rhetorical relations between items. Rhetorical relations are categorized into two types; one represents a logical relation such as resultative conjunction and contradictory conjunction, the

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 790–797, 2008.



| subject | Approach |
|---|---|
| topic | Logical frames are introduced in order to be free from physical constraints of traditional presentation tools. |
| content | Semantic block is divided into some slides because of physical constraints.<br>↓<br>With traditional presentation tools, it is difficult to estimate the relations between different slides.<br>↕<br>Logical frames are introduced in order to be free from physical constraints of traditional presentation tools. |

**Fig. 1.** Structure of a logical frame

other represents a sequential relation. In composing a presentation scenario, it is important for a presenter to consider the relations between topics or items. This is because audiences will not understand the presentation unless they find out the consistent discussion points in its story context.

In order to deal with the problems described above, we address a framework of supporting presentation scenario composition from a viewpoint of intention externalization. In this paper, intention is treated as the presentation strategy of a presenter, and externalization of intention means expressing the procedure of a presentation so that a computer can handle it. This intention is represented as a network structure in which items are related to each other rhetorically. We call this network structure a key phrase channel. By allowing a presenter to specify the relations between items in a presentation scenario, it becomes possible to capture and handle a presenter's intention on the procedure of a presentation.

This paper is organized as follows. Our approach to support presentation scenario composition is described in Section 2. Then, a presentation scenario and its composition process is modeled in Section 3. The implementation of a scenario composition tool is described in Section 4. The possibility of handling a presenter's intention is discussed in Section 5. Finally, our conclusion and future work are presented in Section 6.

## 2 Approach

In traditional presentation tools, the contents of a presentation are divided into blocks according to the physical constraints such as the sizes of slides or displays. However, it is natural that the contents should be divided into semantic

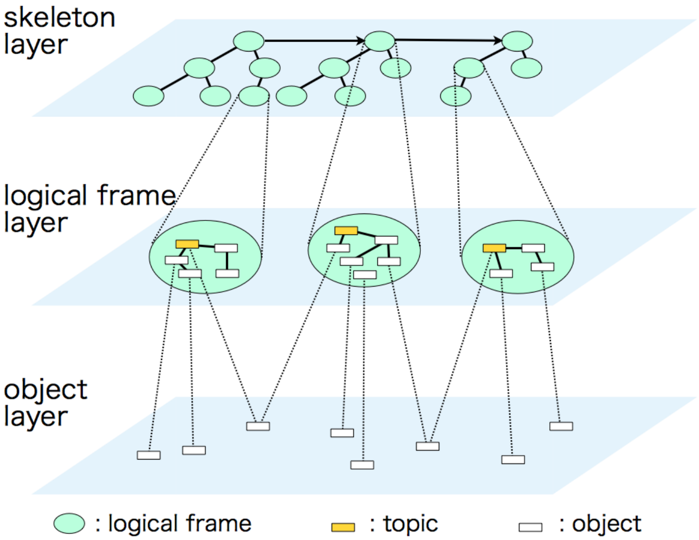


**Fig. 2.** Three-layered model of presentation scenario

blocks of subjects. In order to make it possible to prepare a presentation without considering physical constraints, we introduce a concept of logical frame.

Figure 1 illustrates a structure of a logical frame. A logical frame represents a semantic block of one subject composed of a topic and objects. A presentation scenario is represented as a skeleton structure of logical frames. Logical frames are free from the physical constraints such as the sizes of slides or displays. In each logical frame, idea fragments externalized as character strings or figures, etc. are organized. In the organization process, the items are rhetorically related. Through this process, a presenter is able to externalize his/her presentation intention.

## 3 Presentation Scenario Model

### 3.1 Formulation of Presentation Scenario

The presentation scenario model is illustrated in Figure 2. In our framework, a presentation scenario is represented as a structure with three layers: a skeleton layer, a logical frame layer, and an object layer. In Figure 2, green circles in a skeleton layer and a logical frame layer represent logical frames. Rectangles in a logical frame layer and an object layer represent objects. In our model, a presentation scenario is formulated as follows:

$$\mathrm{Scenario} = (\mathrm{Skeleton}, \mathrm{LogicalFrames}, \mathrm{Objects})$$

A skeleton is a graph structure which represents the relations between logical frames. The data structure in a skeleton layer is formulated as follows:

$$\begin{aligned}\mathrm{Skeleton} &= (\mathrm{V}_S, \mathrm{E}_S)\\ \mathrm{V}_S &= \{f \mid f \in \mathrm{LogicalFrames}\}\\ \mathrm{E}_S &= \{(f_{from}, f_{to}, type) \mid f_{from}, f_{to} \in \mathrm{V}_S\}\end{aligned}$$

In this formulation, $f_{from}$ and $f_{to}$ are references to logical frames in a logical frame layer. *Type* represents a type of relation between two different logical frames.

In a logical frame layer, a set of logical frames is stored. A logical frame is composed of a subject, a topic, and a graph structure. A topic is a string which represents a summary in each logical frame. The topic is assumed to be expressed in one or two sentences. A graph structure represents the relations between objects. The data structure in a logical frame layer is formulated as follows:

$$\begin{aligned}\mathrm{LogicalFrames} &= \{f\}\\ f &= (id, subject, topic, \mathrm{O}, \mathrm{L})\\ \mathrm{O} &= \{o \mid o \in \mathrm{Objs}(f)\}\\ \mathrm{L} &= \{(o_{from}, o_{to}, type) \mid o_{from}, o_{to} \in \mathrm{O}\}\end{aligned}$$

In each logical frame, *id* is its identifier, and *subject* and *topic* are character strings which represent its subject and its summary, respectively. In Figure 2, a topic is represented as an orange rectangle. $O$ is a set of references to objects in an object layer. $\mathrm{Objs}(f)$ is a set of objects included in the logical frame $f$. $L$ is a set of relations between objects included in the logical frame $f$. *Type* represents a type of relations between two different objects.

In an object layer, objects such as character strings, tables, or figures are stored. The data structure in this layer is formulated as follows:

$$\begin{aligned}\mathrm{Objects} &= \{o\}\\ o &= (id, entity, type)\end{aligned}$$

In this formulation, *id* is an identifier of object. *Entity* and *type* are the object itself and its type specified by *id*, respectively.

### 3.2 Scenario Composition Process

The composition process of presentation scenario is divided into two phases: logical frame composition and skeleton composition.

In composing a logical frame, operations in each logical frame are carried out as follows:

- Create a new logical frame.
- Determine *subject* and *topic*.

**Table 1.** Types of relations between objects in a logical frame layer

| Category | Subcategory |
|---|---|
| resultative conjunction | cause-effect<br>if-then |
| contradictory conjunction | opposition<br>concession |
| elaboration | overview-detail<br>general-specific |
| others | supplement<br>parallel<br>others |

– Include/Exclude objects.
– Relate one object to another in the same logical frame.

In the operations listed above, relating objects to each other is realized by specifying the type of a relation between two objects. We categorized the relation types into four groups: resultative conjunction, contradictory conjunction, elaboration, and others. Based on these categories, we classified the types of relations between objects as shown in Table 1.

In composing a skeleton, logical frames are related to each other. The subjects which are semantically related to each other share some concepts. These concepts should be included in the same objects. Therefore, in our framework two logical frames are related if they share one or more objects.

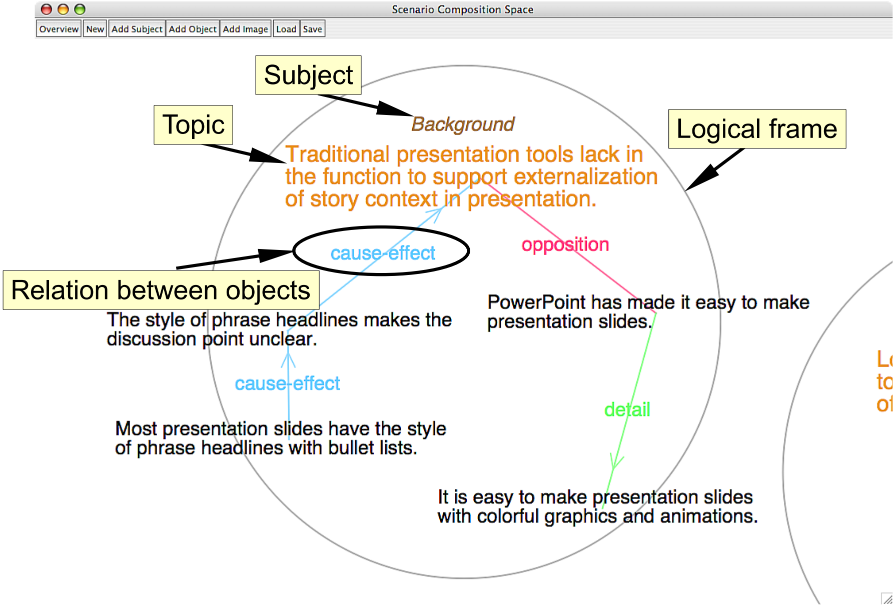
As we discussed in Section 2, a logical frame represents a semantic block such as a chapter, a section, etc. In books or research papers, there are two types of relations between chapters or sections: a sequential relation (previous-next) and a hierarchical relation (section-subsection). Therefore, in our framework the type of a relation between logical frames is either sequential or hierarchical.

## 4 Scenario Composition Interface

Based on the framework described in Sections 2 and 3, we have implemented an interface for scenario composition. In composing a presentation scenario, it is necessary for a presenter to consider the relations or consistency both between logical frames and within one logical frame. Therefore, it is effective to offer a function to change the working view of a presentation scenario. In order to achieve the function, we introduce the technique of zooming and panning in a scenario composition interface.

In implementing an interface for scenario composition, Piccolo[6] is used. Piccolo is a Java library which provides graphical user interfaces with functions including zooming and panning. CounterPoint[5] is one of the tools implemented with Piccolo.

The snapshots of our interface is shown in Figures 3 and 4. In these figures, the area inside a gray circle corresponds to a logical frame. The brown



Traditional presentation tools lack in the function to support externalization of story context in presentation.

PowerPoint has made it easy to make presentation slides.

The style of phrase headlines makes the discussion point unclear.

Most presentation slides have the style of phrase headlines with bullet lists.

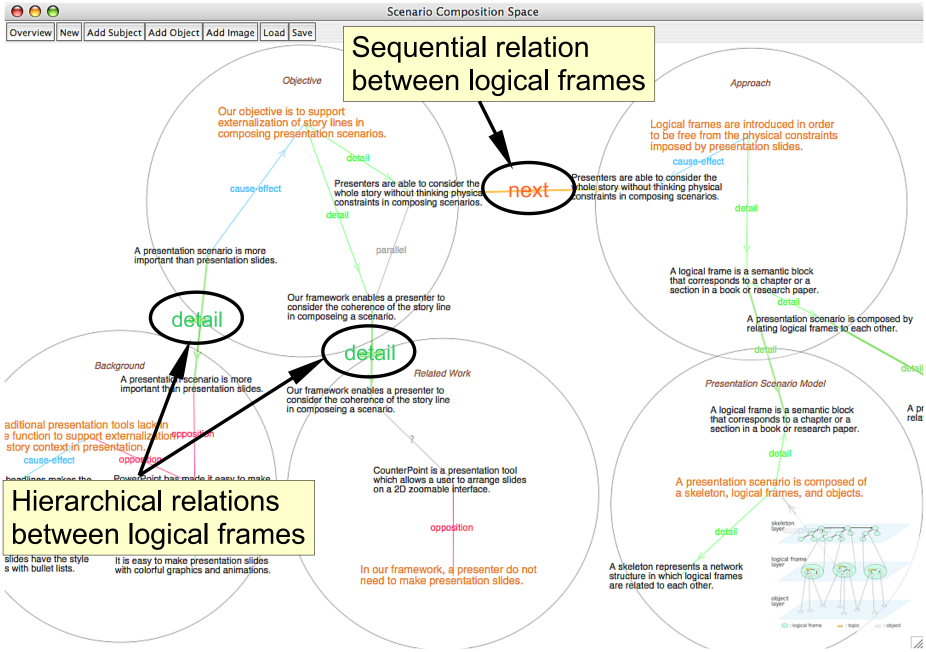
It is easy to make presentation slides with colorful graphics and animations.

**Fig. 3.** Local view of scenario composition interface

text and the orange text represent a subject and a topic of a logical frame, respectively. Relations between objects are represented as links with captions. In Figure 3, for example, a text object "PowerPoint has made it easy to make presentation slides." is related to a text object “It is easy to make presentation slides with colorful graphics and animations.” by the relation “detail”. In the local view shown in Figure 3, a presenter is able to see the detail of a specific logical frame. In the global view shown in Figure 4, a presenter is able to see the whole presentation. In Figure 4, the logical frame “Objective” (the one at the upper left) is related to the two logical frames “Background”, and “Related Work” (the ones at the lower left) with the hierarchical relations. On the other hand, the logical frame “Objective” is related to the logical frame “Approach” (the one at the upper right) with the sequential relation.

In addition to zooming functions, our interface provides a presenter with the operational functions as follows:

- Add a logical frame. This operation is done by specifying a subject and a topic using a dialog box.
- Add an object. This operation is done by specifying an object using a dialog box.
- Relate an object to another one in a logical frame. This is done by dragging and dropping an object near the one that a presenter wants to relate to. After that, the type of the relation can be specified using a dialog box.



Our objective is to support externalization of story lines in composing presentation scenarios.

Presenters are able to consider the whole story without thinking physica constraints in composing scenarios.

A presentation scenario is more important than presentation slides.

Our framework enables a presenter to consider the coherence of the story line in composeing a scenario.

Logical frames are introduced in order to be free from the physical constraints imposed by presentation slides.

Presenters are able to consider the whole story without thinking physical constraints in composing scenarios.

A logical frame is a semantic block that corresponds to a chapter or a section in a book or research paper.

A presentation scenario is composed by relating logical frames to each other.

A presentation scenario is more important than presentation slides.

aditional presentation tools lack in e function to support externalization story context in presentation.

headlines makes the

PowerPoint has made it easy to make

slides have the style s with bullet lists.

It is easy to make presentation slides with colorful graphics and animations.

Our framework enables a presenter to consider the coherence of the story line in composeing a scenario.

CounterPoint is a presentation tool which allows a user to arrange slides on a 2D zoomable interface.

In our framework, a presenter do not need to make presentation slides.

A logical frame is a semantic block that corresponds to a chapter or a section in a book or research paper.

A presentation scenario is composed of a skeleton, logical frames, and objects.

A skeleton represents a network structure in which logical frames are related to each other.

A pr rela

**Fig. 4.** Global view of scenario composition interface

– Relate a logical frame to another one. This is done by dragging and dropping an object in one logical frame into the other one that a presenter wants to relate to. After that, the type of the relation between the two logical frames and the type of the relation between the objects are specified using a dialog box.

Since our interface provides the functions described above, it is possible for presenters to compose scenarios through intuitive manipulation on logical frames and objects. Also, presenters can specify the rhetorical relations between objects, while they can not do that using traditional presentation tools. Moreover, zooming and panning functions enables presenters to look at multiple logical frames at the same time. Therefore, our interface helps them to consider the coherency in their story line of presentation scenarios.

## 5 Discussion

In this section, we discuss the possibility of handling intentions of presenter in terms of a presentation strategy. A presentation strategy is a template of explanation. In presentations, presenters adopt many strategies according to what they talk about. Schank modeled a human memory with a semantic structure called a script[7]. Script is a stereotypical sequence of scenes in a particular event such as taking a meal in a restaurant. Since a presentation is a sequence of explanations in a particular subject, it is possible to represent a presentation scenario in the form of scripts. These scripts correspond to the presentation strategies.

Using the presentation scenario composed with our interface, it is possible to extract the presentation strategy from presentation scenarios. A presentation strategy is considered as frequent patterns in presentation scenarios. In our framework, logical frames include a network structure of objects. Therefore, presentation strategies can be extracted by finding frequent subnetworks in presentation scenarios.

## 6 Conclusion

In this paper, we proposed a framework of supporting externalization of the intentions on a presentation. By introducing a concept of logical frame, we built a presentation scenario model. Based on this model, we implemented a scenario composition interface. The interface has two features. One is to allow a presenter to externalize the rhetorical relation between items. The other is to allow a presenter to change the working view smoothly through the functions of zooming and panning.

One of our future works is to evaluate the effectiveness of our framework through the use study. The effectiveness must be evaluated from the viewpoints of usability, availability for thinking support, and possibility of reuse. Another future work is to devise a mechanism to generate presentation materials according to the presentation scenario. In order to achieve an interactive presentation in which a presenter can control the items to be presented dynamically, we have to consider the method for mapping a presentation scenario onto presentation materials.

## References

1. Apple: Keynote, http://www.apple.com/iwork/keynote/
2. Microsoft: PowerPoint, http://office.microsoft.com/en-gb/powerpoint/
3. Alley, M., Schreiber, M., Muffo, J.: Pilot Testing of a New Design for Presentation Slides to Teach Science and Engineering. In: Proc. of ASEE/IEEE FIE 2005, S3G-7–12 (2005)
4. Tufte, E.R.: The Cognitive Style of PowerPoint. Graphics Press (2004)
5. Good, L., Bederson, B.: Counterpoint: Creating Jazzy Interactive Presentations. Technical Report HCIL-2001 (2001)
6. Bederson, B., Grosjean, J., Meyer, J.: Toolkit Design for Interactive Structured Graphics. IEEE Transactions on Software Engineering 30(8), 535–546 (2004)
7. Schank, R.C., Abelson, R.P.: Scripts, Plans, Goals and Understanding. Lawrence Erlbaum Associates, Mahwah (1977)
8. Schank, R.C.: Tell me a Story. Charles Scribner (1990)

# X-Web: A Data Model for Managing Personal Contents Based on User Experiences

Taketoshi Ushiama[1] and Toyohide Watanabe[2]

[1] Faculity of Design, Kyushu University,
4-9-1 Shiobaru, Minami-ku, Fukuoka 815-8540, Japan
ushiama@design.kyushu-u.ac.jp
[2] Graduate School of Information Science, Nagoya University,
Furo-cho, Chikuka-ku, Nagoya 464-8603, Japan
watanabe@is.nagoya-u.ac.jp

**Abstract.** In this paper, we propose the X-Web (eXperience-Web) data model, which supports users to manage various types of personal contents in a unified manner based on their contextual information. The X-Web data model consists of three kinds of modeling units: contents, experiences, and persons, and the context of personal contents are expressed as experiences. Units are classified into containers, and a ranking function is defined between two containers. Combinations of ranking functions can represent various requirements for searching and recommending personal contents.

## 1 Introduction

Recently, personal contents that an individual person must manage increase rapidly, and researches on personal content management have been activated. However, most of conventional techniques cannot support a user to manage various types of contents in a unified manner. Moreover, a user cannot manage contents effectively with only objective meta-data, because contextual information is necessary to manage them efficiently. However, general frameworks for supporting such contextual information are not proposed.

This paper introduces the X-Web (eXperience-Web) data model, which supports users to manage their personal contents based on their experiences in the real and virtual world. The X-Web data model provide data structure on the Internet for organize various types of contents based on personal experiences in a uniform way. The X-Web data model enables users to search and to recommend variety types of personal contents based on their contextual information. Fig. 1 illustrates an overview of the system.

On the X-Web system, personal contents managed by users are gathered and allowed to be accessed. It is an important issue how to share personal contents of many users in a distribute environment and how to protect privacy of users, but we will not address these issues in this paper.

This paper contains followings: Section 2 describes a basic idea for managing personal contents by means of user experiences. Section 3 define the X-Web

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 798–805, 2008.

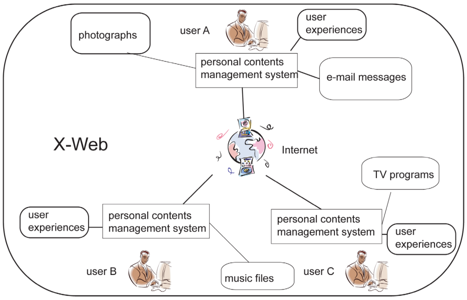


**Fig. 1.** An overview of the X-Web system

data model. Section 4 describes how to rank personal contents according to a requirement of a user. Section 5 conclude this paper and shows some future works.

## 2 Personal Contents and User Experiences

### 2.1 Categorization of User Experiences Based on Rolls of Personal Contents

Personal contents are the contents that users manage and utilize personally. Roles of personal contents are categorized into the following two types:

1. recording some situations and information about experiences of a user, and
2. extracting information and knowledge or enjoying by watching, reading, and listing.

Personal contents can be related to a user who manages them on the bases of his/her experiences. According to the varieties of roles of personal contents, the user experiences concerned with personal contents can be categorized into two categories: (1) active experiences and (2) passive experiences. An active experience is a real word event that internalized into one or more contents. On the other hand, passive experience is an experience in which a user enjoys by reading, watching, and listening of a content object. A passive experience is an experience that is internalized into a user from a content object.

The X-Web data model stores active experiences, which are recorded in contents. It also stores passive experiences, which are activities of utilizing contents. The stored experiences are used for retrieving and recommending personal contents. Representative examples of active experiences are shooting photos, sending/receiving e-mails. Representative examples of passive experiences are reading web pages, watching TV programs, listing songs.

### 2.2 Utilization of Experience for Managing Personal Contents

Personal contents can be categorized into private contents and public contents.

Many of personal contents recorded in an active experience are private contents. Photos taken by digital cameras and email messages are representative examples of private contents. It is effective for managing private contents to retrospective navigation and associative search based on personal experiences of a user. Storing active experience supports personal contents searching based on various features in different contents and context based searching based on attributes of experiences.

Public contents are the contents that are used for obtaining information and enjoying by public users. Most of commercial contents can be viewed as public contents. Most of public contents such as TV programs, movies, songs and so on can be associated with passive experiences. Some of private contents can be associated with passive experiences, where a user watches a photo taken by the user. It is important for public contents to support recommendation based on personal profile of a user. [1]. Storing passive experiences can be used for recommending contents by means of estimating user's preference.

## 3 X-Web Data Model

In this section define the X-Web data model. Firstly, we introduce a model for search and recommendation, and then we define the data structure and operations of the model.

The X-Web data model provides a unified framework for search and recommendation. The framework is that search and recommendation are achieved by means of weight propagations among conceptual units. The weight values are assigned according to a query of a user.

In the X-Web data model, modeling primitives are called *units*. The set of all units is represented by $\mathcal{U}$. A subset of $\mathcal{U}$ is named *container* and is represented by $C_i$.

$\langle C_1, \cdots, C_n \rangle$ represents a sequence of $n$ containers $C_1, \cdots, C_n$ where $u_j^i$ represents a unit in a container $C_i$, the weight values from $u_j^i$ to $u_k^{i+1}$ is represented by $w_{jk}^i$. The weight value represents the degree of adaptation for a user's requirement.

The tail container $C_n$ is the target set for searching and recommending, and the head container represents conditions for a user's query. The initial value is given for the head container. The initial value for a unit $u_j^0$ is represented by $w_j^0$.

When a user executes a query, the system calculates the weight value of every element in the last container $U_n$ by weight propagation, after the container sequence and the initial weight value is specified by the user. The final weight value of $u_j^n \in U_n$ is represented by $w_j^n$. In order to calculate $w_j^n$ the following formula is used.

$$w_l^n = \sum_i \sum_j \cdots \sum_k w_i^0 w_{ij}^1 \cdots w_{kl}^{n-1} \quad (1)$$

The last weight values are viewed as the degree of adaptation. The larger the weight values of a unit are, the unit is the more adaptable for a user requirement. Fig. 2 shows an overview of data structure for search and recommendation of

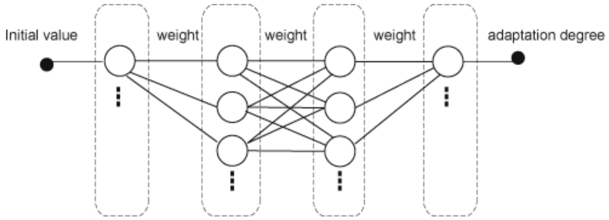


**Fig. 2.** The basic model for searching and recommending

the X-Web data model. A circle represents a unit, and a rectangle with dashed line represents a container.

The X-Web data model provides search and recommendation functions using the above data structure. The following steps are required for the functions:

1. specifying a sequence of containers,
2. assigning the weight value to each link between two units in a pair of neighboring containers.
3. assigning the initial weight value to each units in the head container of the sequence, and
4. calculating the weight value of each units in the last container according to the formula (1).

**Function for Search of the X-Web Data Model.** This section describes how to implement searching function based on the basic model. A query is represented by the weight value of units in the head container of a container sequence. The last container represents the target set for searching. The weight value of a unit in the last container represents the degree of sufficiency as a result of the query.

Fig. 3 (1) shows an example of a traditional document search. $C_1$ is the container for the set of terms specified by a user as a query. $C_2$ is the container for the set of index terms. $C_3$ is the container for the set of documents which are search targets.

Fig. 3 (2) shows an example of searching private photographs of a user using his/her e-mail messages. $C_1$ is the container for query terms of a user. $C_2$ is the container for e-mail messages of the user. $C_3$ is the container for temporal descriptions in the e-mail messages $C_2$. $C_4$ is the container for photographs which are searching targets.

**Function for Recommendation of the X-Web Data Model.** This section describes how to implement a recommendation function on the X-Web data model.

Many varieties of recommendation techniques have been proposed up to now. Collaborative filtering[2] is the one of the most popular recommendation techniques. The collaborative filtering technique enables to recommend unknown contents to a user based on evaluation of other users.

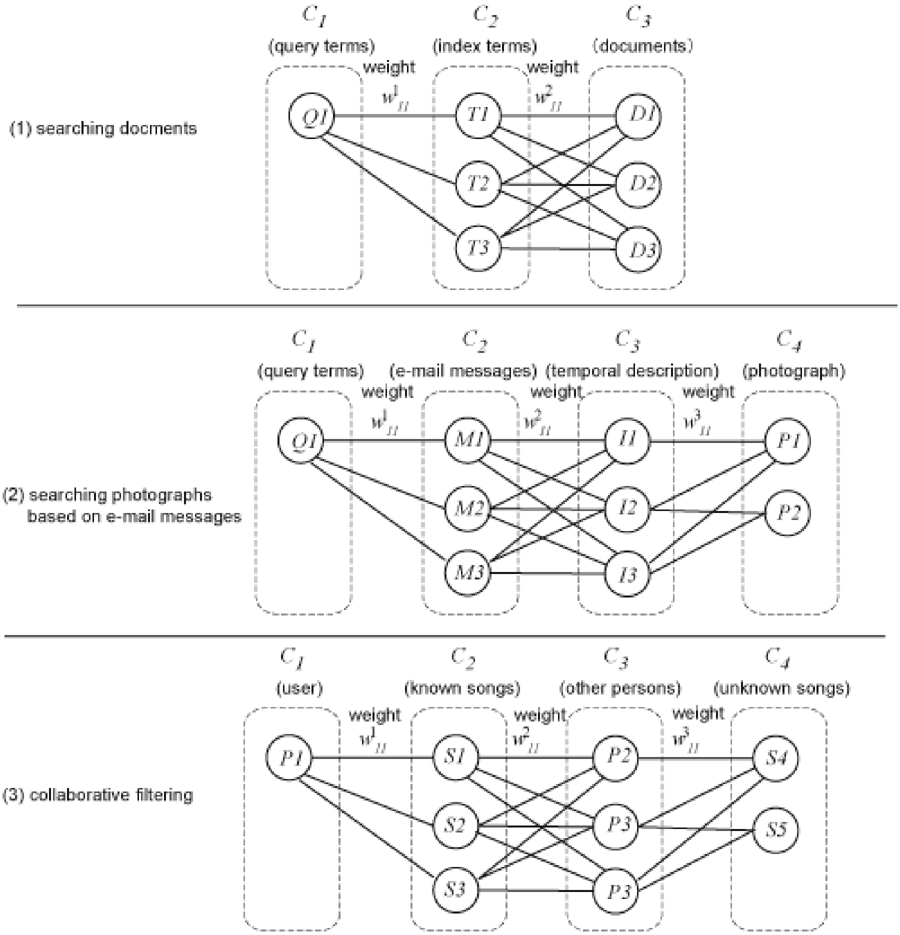


**Fig. 3.** Three examples of search and recommendation

Fig. 3 (3) shows an example of a song recommendation like collaborative filtering. $C_1$ is the container for a user $P1$ who take song recommendations. $C_2$ is the container for songs which have been evaluated by the user $P1$. $C_3$ is the container for other users $P2$, $P3$, and $P4$ who are used for criteria of recommendations. $C_4$ is the container for unknown songs $S4$ and $S5$ which are candidate for recommendation. The weight value of a song in $C_4$ is treated as the degree for recommending. This scheme supports recommendation results which are similar to GroupLens[3].

### 3.1 Data Structure

The following shows formal definitions of modeling elements of X-Web.

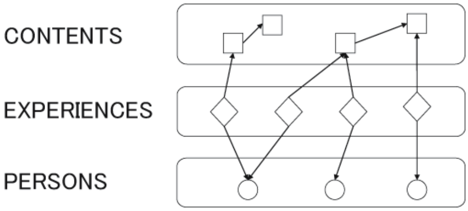


**Fig. 4.** Data structure

**Unit.** Modeling primitives of the X-Web model are units. Units are categorized into four types: contents, experiences, persons and concepts.

A content unit is defined as $c = (cid, data, A_c)$ where $cid$ is a content identifier, $data$ is a reference to the content, $A_c$ is an attribute.An attribute is defined as a set of pairs of attribute name $a_i$ and value $v_i$, which are represented by $A_c = \{(a_1, v_1), \cdots, (a_n, v_n)\}$. A attribute value can be a number or character string.

An experience is defined as $e = (eid, c, p, t, A_e)$, where $eid$ is an identifier of experiment, $c$ is a reference to a content unit, $p$ is a reference to a person unit, $t$ is a time stamp $A_e$ is a set of attributes. A person unit is defined as $p = (pid, A_p)$ where $pid$ is an identifier of a person and $A_p$ is a set of attributes.

Fig. 4 illustrates relationships between three types of units. In this figure a content unit is represented as a rectangle, a person unit is represented as an oval and an experience unit is represented as a lozenge. A reference from an experience unit to a content unit or a person unit is represented as an arrow.

**Container.** Containers are used for representing sets of units. A container is defined $C = (name, \{u_1, \cdots, u_n\})$ where $name$ is a name of container and $\{u_1, \cdots, u_n\}$ is a set of units.

The container which contains all person units is represented as “PERSONS”, the container which contains all experience units is represented as “EXPERIENCES” and the container which contains all content units is represented as “CONTENTS”. A rectangle with rounded corner is a container in Fig. 4.

There are two types of approaches for specifying elements of a container. One is to specify elements in a container explicitly, and the other is to specify conditions which elements in a container must satisfy. When a user specifies elements of a container with conditions, a container including units which satisfy the conditions $cond$ is denoted by $C[cond]$.

Search and Recommendation on the X-Web data model are realized by means of ranking networks. A query is represented by a ranking network. In order to specify ranking network, it is required for a user to specify a sequence of containers and weight values among units in the containers. Weight values among units are represented by a ranking function. A ranking function is defined as $(N, C_S, C_T, W)$ where $N$ is a function name, $C_S$ is a source container, $C_T$ is a target container and $W$ is a set of weight values from an element in $C_S$ to

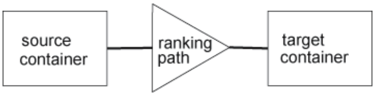


**Fig. 5.** An overview of ranking function

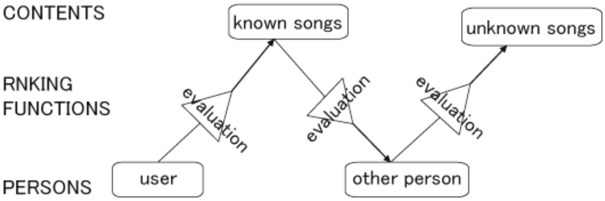


**Fig. 6.** An example of a ranking path

an element in $C_T$. Multiple ranking functions are assigned to a single pair of containers. Fig. 5 illustrates the structure of a ranking function.

**Query Specification with Ranking Path.** On the X-Web data model stored data are represented as a directed graph. A weight propagation network can be organized where a path on the directed graph. This means that a query can be represented by a path on the directed graph because a query is specified by a weight propagation network. A path which represents a query is named as a ranking path. A ranking path is defined as $(\langle r_1, \cdots, r_n \rangle, W_0)$ where $r_i$ is a ranking function and $W_0$ is a set of initial values for the units in the head container. Fig. 6 shows an example of a ranking path. This example provides a collaborative filtering on songs.

## 4 Assigning Weight Values on Ranking Functions

### 4.1 Two Approaches for Assigning Weight Values

Techniques for assigning weight values on a ranking function are classified into two types. One is the manual assignment approach and the other is the automatic assignment approach. Manual assignment techniques enable to assign various types of evaluation as weight values, but they require a user to do some tasks. On the other hand, automatic assignment techniques utilize features of units for assigning weight values. For example, TF*IDF techniques use term frequency and document frequency for assigning weight to document contents and the PageRank algorithm[4] utilize the link structure for assigning weight values to web page contents. Today many tools and APIs for analyzing contents have been provided and they can be adapted for assigning weight values on the X-Web data model. For example, a ranking function for web page units can be implemented with the Google search API.

### 4.2 Assigning Weight Values Based on User Experiences

It can be considered that experience units represent context about content units. Context information can be utilized for ranking content units. In order to calculate weight values, a user specify formulas using attributes and aggregate functions. An attribute is described by dot expression. For example the expression "Play.playtime" represents the attribute "playtime", which shows how long a user has listened a song, of an experience unit in the category "Play". The X-Web data model supports basic mathematical operators and typical aggregate functions such as COUNT, MIN, MAX, SUM, and AVG. iPod supports a function for playing song files in the order of how many times a user play a song file. This function is designed based on an assumption that the number of playing a song reflects how much a user prefers the song. A similar function can be implemented on the X-Web data model using the following expression: $(\mathrm{Users}, \mathrm{Songs}, \mathrm{SUM}[\mathrm{Play.playtime}])$. This function provides ranking songs based on total listening time of a user.

## 5 Conclusion

This paper introduces the X-Web data model, which provides a user to manage various kinds of personal contents, and give a formal definition of its data structure. The X-Web data model enables a user to search and recommend by means of the same model named as weight propagation network, and to rank personal contents based on behaviors of users. We plan to implement a system prototype and to evaluate the X-Web data model based on experimental results.

## References

1. Schafer, J.B., Konstan, J.A., Riedl, J.: E-commerce recommendation applications. Data Mining and Knowledge Discovery 5(1/2), 115–153 (2001)
2. Shardanand, U., Maes, P.: Social information filtering: Algorithms for automating "word of mouth". In: Proceedings of ACM CHI 1995 Conference on Human Factors in Computing Systems, vol. 1, pp. 210–217 (1995)
3. Resnick, P., Iacovou, N., Suchak, M., Bergstorm, P., Riedl, J.: GroupLens: An Open Architecture for Collaborative Filtering of Netnews. In: Proceedings of ACM 1994 Conference on Computer Supported Cooperative Work, Chapel Hill, North Carolina, pp. 175–186. ACM, New York (1994)
4. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project (1998)

# Understanding Creativity-Technique Based Problem Solving Processes

Florian Forster and Michele Brocco

TU Muenchen
forster@in.tum.de, brocco@in.tum.de

**Abstract.** Computer-supported creativity techniques can help people finding creative solutions for their problems. However, real-life creative processes demand a high level of flexibility of the support systems, which are normally tailored for one specific creativity technique only. We present a model for creativity-technique based problem solving processes that incorporates a variety of creativity techniques and can be a promising starting point towards more flexible creativity support systems in the future.

## 1 Introduction

In the 1960s Alex Osborn proposed to enhance the outcome of a group searching for creative solutions to a given problem by following certain guidelines: deferring judgement, preferring quantity over quality and generally avoiding any kind of censorship [1]. In the following, literally hundreds of these creativity techniques were proposed and researched, providing findings that most of them can improve creative problem solving. Starting in the early 1990s, scientists have been trying to find out how computer systems could help improving the creative problem solving process. Empirical results show that groups using these creativity support systems (CSS) generate more ideas and more creative ideas than control groups working without computer support [2].

## 2 Static Process Models and Dynamic Processes

From the great number of existing creativity techniques, only a small fraction was adopted for computer use and researched, e.g. the brainstorming technique, brainwriting-variations, morphological analysis and mindmapping. The applied tools are tailored for one creativity technique. Hence, the underlying process model is static and highly specific. In contrast, real-life creative problem solving processes show dynamic elements:

– Problem solving processes may be iterated multiple times until a desired result is accomplished. The processes can even be nested.
– Not all creativity techniques are useful for every problem or in every context. One should choose the most suitable technique for a given situation.

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 806–813, 2008.

- It can be useful to apply multiple creativity techniques for a given problem to profit from the advantages of each technique.

In this paper, we want to present an approach which can help developers of support systems for creative problem solving to overcome this issues.

## 3 Understanding the Creative Process

We propose a model of computer-supported creative problem solving processes that can help to understand dynamic creative processes better. The model was developed in two steps:

1. A creative problem solving process is a special form of human thinking process. We discuss at which points computers generally can support such processes and propose a simple Input/Output-Model as a starting point.
2. A creative problem solving process is characterised at the most abstract level as the sequence of one generation (divergent) and one evaluation (convergent) process phases. We discuss the implications for our model and extend it accordingly.

Creativity techniques are definitions for creative problem solving processes that have shown to be more effective than other approaches. We describe the strong similarities between creativity techniques and computer support for creative problem solving. A lot of techniques have been proposed in literature, and at first sight it appears contradictory or even impossible to integrate this variety in a process modell that promises generality. However, our analysis of more than 25 creativity techniques for idea generation and idea evaluation showed that the techniques are based on a surprisingly small number of functional patterns. We describe these patterns in detail and show that they can easily be incorporated in our model as they can be classified either as input- or output-transformations. These findings imply that the proposed model can be a fruitful starting point to build more comprehensive and flexible creative support systems.

## 4 A Flexible Process Model of Creative Problem Solving

As stated above, our goal was to develop a flexible model of the creative problem solving process that can be used to build more effective creativity support systems. In the following, we discuss the development of our model in detail:

### 4.1 Computer Support of Thinking Processes

Thinking processes are cognitive processes than run in the peoples' brains. These cognitive processes do not follow strict logical rules and are so complex that they can neither be influenced predictably nor can they be imitated by computer systems. This is a noteable difference to most other domains of computer support,

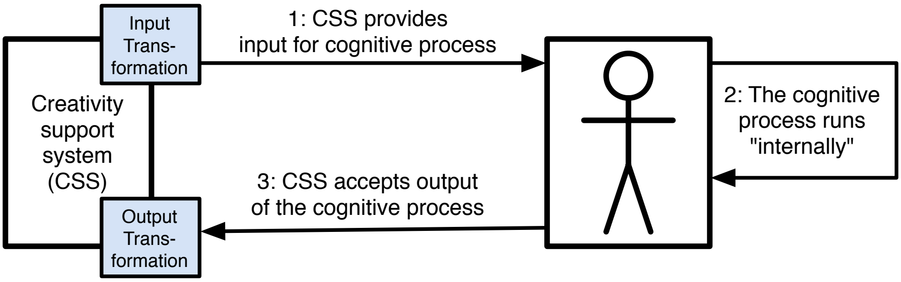

Input Transformation
1: CSS provides input for cognitive process
Creativity support system (CSS)
2: The cognitive process runs "internally"
3: CSS accepts output of the cognitive process
Output Transformation

**Fig. 1.** Model of the computer-supported cognitive process

where either the computer can solve the task by itself (searching data, sorting, filtering) or can at least ease the task by partially ”thinking like the user” (recommender systems, semantic web). Therefore, we propose to regard the creative thinking process simply as a black box that produces thoughts from input impulses. One can think of a computer-supported thinking process as a composed process running in two distributed components: the computer and the human (see fig. 1). The computer kicks off the process by displaying arbitrary information to the user. This information activates cognitive processes in the human brain. After this process is finished, the human can enter the results (i.e. his thoughts) back to the support system, where they can be processed further. Thus, computer support systems trying to affect creative cognitive processes are limited to work on the interfaces of the human-computer-interaction (depicted as blue filled squares in fig. 1). At the interfaces, the systems can *transform* the information of the thinking process, i.e. add supplemental information, modify existing information or hide existing information.

1. Transforming the input for the human thinking process (Step 1 in fig. 1): From the perspective of a support system this means displaying any type of information to the user with the goal to bias his internal thinking process.
2. Transforming the output of the human thinking process (Step 3 in fig. 1): After the thinking process delivered some results (thoughts), the user has to communicate these results back to the support system. The results of the overall process can therefore be influenced by offering certain possibilities or imposing limitations at this point.

### 4.2 Computer-Support of Creative Problem Solving Processes

A creative problem solving process is a special form of thinking process with the goal to find creative *ideas* for a given *problem*. The process is commonly described as phase model, and the various proposed process models differ in their level of abstraction as well as in their boundaries. A general pattern that can be found in

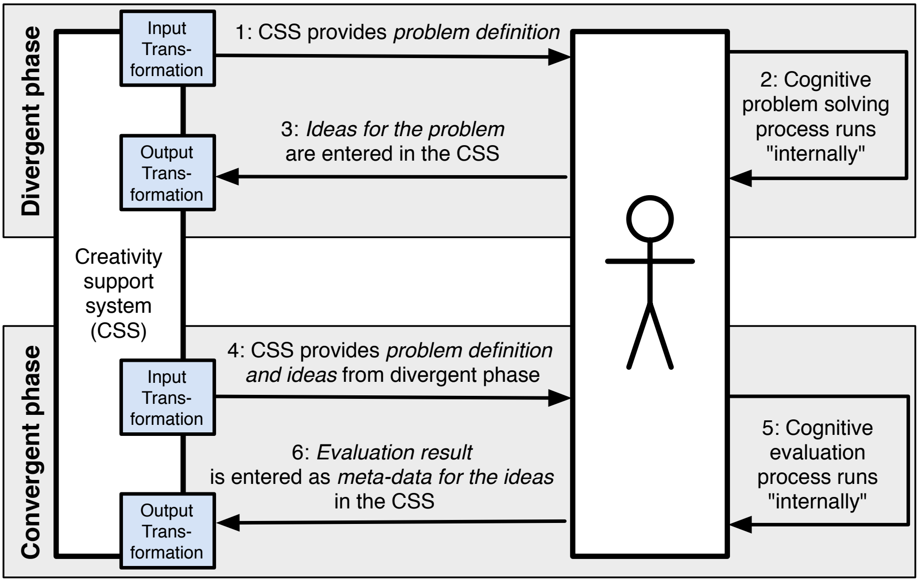

Divergent phase
Input Transformation
1: CSS provides *problem definition*
2: Cognitive problem solving process runs "internally"
3: *Ideas for the problem* are entered in the CSS
Output Transformation
Creativity support system (CSS)
Convergent phase
Input Transformation
4: CSS provides *problem definition and ideas* from divergent phase
5: Cognitive evaluation process runs "internally"
6: *Evaluation result* is entered as *meta-data for the ideas* in the CSS
Output Transformation

**Fig. 2.** Model of the computer-supported creative problem solving process

most of the process model descriptions, as well as in many creativity techniques, is the sequence of two distinctive phases [3]: A divergent phase, where ideas are generated and collected, followed by a convergent phase, where the ideas are discussed, evaluated and finally discarded or selected.

Adding this concepts to our model of cognitive process support results in the refined model depicted in fig. 2. The computer-supported creative problem solving process is modeled as two consecutive computer-supported cognitive processes. The type of information that is exchanged between the computer and the human is clearer defined due to the definition of the creative problem solving process. In the divergent phase, the CSS has to provide at least a definition of the problem, and the response by the human is to be interpreted as idea for this problem. During the convergent phase, the CSS has to set the context for the user's evaluation process by displaying both the problem and the ideas from the previous divergent phase. The user's response during this phase has to be interpreted not as ideas, but as data about the ideas (*meta-data*).

Given these considerations, designing an effective support system for creative problem solving can be regarded as the search for appropriate transformations for the input and output of the human cognitive processes involved in the divergent and the convergent phase. Of course, there are unlimited ways to transform the information provided to the cognitive process or to modify the information that flows back as response. How can we identify the appropriate transformations, hence transformations that will have a positive impact on the result of the creative problem solving process?

### 4.3 Computer-Support of Creativity-Technique Based Problem Solving Processes

From a process point-of-view, a creativity technique is a description or template for creative problem solving processes. The techniques define certain presettings to or constraints on the processes, promising that these specialised forms of creative problem solving processes are effective in practice. This is the same objective that is pursued by creativity support systems. Similar to creativity support systems, creativity techniques are also limited to the same sphere of influence. They can define the input (prepare the context) for the human cognitive process and imply rules on how the mental output should be materialized. Given this strong structural similarity, it becomes obvious that we can learn a lot about building more effective creativity systems by understanding the effectiveness of creativity techniques.

Following this thought, we analyzed a multitude of common creativity techniques from two different sources [4] [5]. Our goal was to find out how exactly each of it specialises the creative problem solving process. Afterwards, we tried to identify repeating patterns in these specialisations. It pointed out that creativity techniques are based on surprisingly few basic patterns. All the described patterns can be considered either as input or output transformations of the underlying cognitive processes, hence they consistently integrate into our model of computer supported creative problem solving processes. Table 1 lists all investigated techniques and the patterns they use. We now want to discuss the patterns in detail:

**Transformation patterns of creativity techniques:**

1. General patterns
   The following patterns were found in both generation and evaluation phases of creativity-technique based processes:
   - *Process nesting*: A large number of the investigated techniques are based on nesting creative processes: Before the actual problem $P$ can be solved, other problems (based on $P$) have to be solved first. The results of these sub-processes are at the end used as input for the actual solving of $P$. For example, the Morphological Analysis works by nesting processes: Before one can solve a a problem $P$ with the Morphological Analysis, one has to first solve a subprocess to identify components of $P$. Then, for each of the components, subprocesses to find appropriate attributes of each componenent have to be solved. Finally, the results of the process are used as starting point to generate ideas for the actual problem. Similar, all of the various checklist-based techniques (like the SCAMPER or the CATWOE technique) rely on the nesting process pattern.
   - *Time limit*: A few of the investigated techniques strictly impose time limits on the participants (like the Brainwriting 6-3-5 technique, which is done in 6 iterations with a 5 minute time limit for each iteration), even though setting a time limit for the phases of the creative process can generally be appropriate, at least due to practical and organizational reasons.

2. Generation phase patterns

   The following patterns were found in generation phases of creativity-technique based processes only:

   - *Stimuli*: A stimulus can be any type of information that is used to impact the cognitive process (see 4.1). For generation phases, stimuli are a very common pattern. The stimuli are often generated in subprocesses by the group itself.
   - *Start ideas*: In contrast to stimuli, start ideas can be used directly to construct new ideas, either by combining the start ideas or by mixing in own ideas. An example for this pattern is the morphological analysis that is based on using results of subprocesses as start ideas for the actual problem solving process.
   - *Idea representation limit*: Some techniques limit the participants in the way they can express ideas. E.g. the Brainsketching technique forces the participants to express their ideas as sketches or the Greeting Cards method defines to use pictures from external sources only.
   - *Idea number limit*: Limiting the number of ideas the participants can contribute intuitionally seems to be counterproductive, but can be necessary to ease group processes like Brainwriting 6-3-5..

3. Evaluation phase patterns

   The following patterns were found in evaluation phases of creativity-technique based processes only:

   - *Criteria*: The criteria pattern is by far the most frequent modification pattern in the evaluation phases of the creative process. In the standard evaluation process the participants assess if the whole idea is a valid solution for the problem. When the criteria pattern is used, each idea is evaluated against a set of criteria instead. The criteria can either be static (the same set of criteria for all creative processes) or dynamic (the criteria are generated during the process). E.g. the advantage-disadvantage-technique defines that the participants have to compile a list of criterions in a subprocess (*process nesting pattern*) first, and afterwards assess for each idea and each criterion if the idea would affect the criterion positively (advantage) or negatively (disadvantage). In contrast, the Castle technique defines three static criteria: acceptability, practicality and originality.
   - *Scenarios*: A scenario is a description of a plausible future. A number of investigated evaluation techniques advise to evaluate the ideas against a set of scenarios to better understand the implications and especially the risks of implementing the idea (e.g. the GAS / Four Futures technique).
   - *Anonymity*: By providing anonymity, negative social group effects can be avoided. Anonymity can be regarded as a transformation of the output of the evaluation process (removing user information).
   - *Scoring*: Scoring (or voting) means assigning points to ideas. In most cases, the ideas are ranked by their total score afterwards. Scoring can be seen as a limitation of the output of the evaluation process, as it restricts the user to numeric values.

Summarizing, or findings show that a creativity support tool that incorporates these patterns as input- or output-transformations of the creative process can be used to support a great variety of different creativity techniques. This indicates that the proposed model is an appropriate model of the creative process that can help to build more flexible creativity support systems.

**Table 1.** Creativity technique transformation patterns

| Technique [Source] | Input transformations | Output transformations |
|---|---|---|
| Analogies [5] [4] | Process Nesting, Stimuli | |
| Assumption Reversals [5] [4] | Process Nesting, Stimuli | |
| Bionics [4] | Process Nesting, Stimuli | |
| Boundary Examination [5] [4] | Process Nesting, Stimuli | |
| Brainwriting 6-3-5 [5] [4] | Process Nesting, Start ideas | Time limit, Idea number limit |
| Brainsketching [5] | Process Nesting, Start Ideas | Time limit, Idea repr. limit |
| CATWOE [5] | Process Nesting, Stimuli | |
| Collective Notebook [5] [4] | Process Nesting, Start ideas | |
| Five W's and H [5] [4] | Process Nesting, Stimuli | |
| Free Association [5] | Process Nesting, Stimuli | |
| Greeting Cards [5] | Process Nesting, Stimuli | Idea representation limit |
| Morphological Analysis [5] | Process Nesting, Start ideas | |
| Problem Reversal [5] [4] | Process Nesting, Stimuli | |
| Progressive Abstraction [4] | Process Nesting, Stimuli | |
| Random Stimuli [5] [4] | Stimuli | |
| SCAMPER [5] | Process Nesting, Stimuli | |
| Advantage-Disadvantage [4] | Process Nesting, Criteria | Scoring |
| Adv., Lmt. and Unique Qual. [5] | Criteria | Scoring |
| Anonymous Voting [5] | | Anonymity |
| BulletProofing [5] | Process Nesting | Scoring |
| Castle Technique [4] | Criteria | Time limit, Scoring |
| Creative Evaluation [4] | Criteria | Scoring |
| FBAS [5] | Scenarios | |
| GAS (Four Futures) [5] | Scenarios | |
| Negative Brainstorming [5] [4] | Criteria | |
| Sticking Dots [5] [4] | Criteria | Scoring |

## 5 Related Work

Quite a few empirical results on using creativity techniques with computer support systems have been published so far. Bostrom and Nagasundaram provide an excellent overview of the studies in this field up to the late 1990s [2]. More recent examples are the work by Janssen et al. [6], who experimented with creativity methods in CSCW envirionments, and the article written by Neupane et al. [7], who investigated which effect on the quality and quantitiy of ideas is attained by changing certain parameters in the brainwriting 6-3-5 technique. We propose to continue working in this direction, conducting similar experiments by varying the transformation patterns we identified. Resnick et al. summarized important general guidelines for designing creativity support tools [8], confirming the importance of flexible solutions. In one of the rare theoretical treatments in the field, Casalini et. al. [9] suggest a formal process model for collaborative problem solving especially for virtual communities in practice, however without addressing creativity aspects.

## 6 Conclusion

We presented a model of the computer-supported creative problem solving process. The model is based theories of supporting cognitive processes in general and creative processes in particular. We identified functional patterns within creativity-technique based processes and showed that these patterns can be interpreted either as input or output-transformations of the process model. Hence, building up on the proposed process model is a promising approach towards more dynamic and flexible creativity support systems.

## References

1. Osborn, A.F.: Applied imagination. Principles and Procedures of Creative Thinking, 3rd edn. Creative Education Foundation (1993) ISBN 0930222733
2. Bostrom, R.P., Nagasundaram, M.: Research in creativity and gss. In: HICSS 1998: Proceedings of the Thirty-First Annual Hawaii International Conference on System Sciences, Washington, DC, USA, vol. 6, p. 391. IEEE Computer Society, Los Alamitos (1998)
3. VanGundy, A.B.: Idea Power. AMACOM (1992)
4. VanGundy, A.B.: Techniques of Structured Problem Solving. Van Nostrand Reinhold (1988)
5. Mycoted: Creativity & innovation, science & technology, http://www.mycoted.com
6. Janssen, D., Schlegel, T., Wissen, M., Ziegler, J.: Metachart: Using creativity methods in a cscw environment. In: Proceedings of the 10th International Conference on Human-Computer Interaction, vol. 2, pp. 939–943 (2003)
7. Neupane, U., Miura, M., Hayama, T., Kunifuji, S.: Qualitative, quantitative evaluation of ideas in brain writing groupware. IEICE Transactions 90-D(10), 1493–1500 (2007)
8. Resnick, M., Myers, B., Nakakoji, K., Shneiderman, B., Pausch, R., Selker, T., Eisenberg, M.: Design principles for tools to support creative thinking. In: NFS Workshop Report on Creativity Support Tools. National Science Foundation (2005)
9. Casalini, M.C., Janowski, T., Estevez, E.: A process model for collaborative problem solving in virtual communities of practice. Technical report, United Nations University - International Institutre for Software Technology (April 2007)

# Visualisation Tool for Cooperative Relations and Its Verification with a Case Study

Takanori Ugai and Kouji Aoyama

Fujitsu Laboratories Limited,
4-1-1 Kamikodanaka Nakaharaku Kawasaki Kanagawa 211-8588, Japan
{ugai,aoyama.kouji}@jp.fujitsu.com

**Abstract.** Many organisations have communication problems such as inter-organisational barriers. However, these kinds of problems are not easy to visualise, and it is even more difficult to derive, implement and assess appropriate measures to deal with them. We developed a tool to visualise the dynamic structure of cooperative relationships between employees in organisations based on questionnaires given to employees of those organisations. This tool is used for visualising barriers between teams and the effects of measures. In this paper we explain some features of this tool and verify its capabilities and effectiveness with a case study. The case study is some field research based on interviews that we conducted in which we applied measures to improve the employees' communication. We collected a set of data about relationships in an organisation with questionnaires before and after implementing the measures. And we compared the observed result produced by the visualisation tool with the result from the field research.

**Keywords:** Knowledge Management, Cooperative Relationship, Social Network Analysis, Visualisation, Case Study.

## 1 Introduction

It is very important to have a company that is growing and whose entire workforce is activated. To have such a company, one of most important disciplines is to collect a lot of useful information and effectively use the knowledge of individuals within that organisation[1].

We have proposed a mathematical model for transferring knowledge with the purpose of making a KM mechanism or system take root in an organisation and obtaining guidelines to make that system functional[2]. We have applied this model to some cases and produced guidelines for appropriate systems to manage knowledge[3].

We have found that many organisations have communication problems such as inter-organisational barriers. Mistakes occurred because of a lack of communication. The efficiency of work does not improve because information and knowledge do not spread in the organisation owing to this lack of communication.

However, this kind of problem is not easy to visualise, and it is even more difficult to derive, implement and assess appropriate measures for it.

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 814–822, 2008.

In many cases, people in the organization realise there is a lack of communication after an accident has occurred, but it is often too late to take measures. Generally, measures to eliminate a lack of communication cannot be expected to produce short-term results. Because organisations change and improve gradually, using indicators such as the decreased number of mistakes or ROI is inappropriate.

We developed a tool to visualise the dynamic structure of cooperative relationships between employees of an organisation based on questionnaires given to employees. This tool visualises not only the organization's static structure but also its dynamic structure. It can clearly visualise barriers in an organisation and the effect of measures which the organisation has taken.

In this paper, we propose a tool which shows a layout of employees as nodes and shows the cooperative relationships between them as edges. It visualises the static structure of cooperative relationships and the dynamic structure of relationships as a sequence of static structures and shows any difference between those static structures. We also describe how this tool works with a case study. The case study is based on our field research on an organisation that develops software. We compare the field research with findings from the tool's visualisation based on a questionnaire given to the employees.

There are such visualisation and analysis methods and tools in existence today[4,5,6]. However, they are general analysis tools, and it is difficult to understand problems which happen in the field from the diagrams which those tools draw. The users have to analyse and interpret the meaning of the diagrams and numbers by making investigations in the field. Our tools target the managers of an organisation. It should show what kinds of communication problems their organisation has in an intuitive way.

Using the results of the field research, we have developed some knowledge management guidelines that are designed to apply the model in a systematic derivation of measures that need to be implemented in the organisation. We show some diagrams made using results of the questionnaire given to employees about their cooperative relationships. And we validate the findings from those diagrams with findings from some field research. This validation demonstrates the effectiveness of the tool.

In chapter 2, we describe the functions of the tool. In chapter 3, we describe the organisation's profile and a preliminary survey. In chapter 4, we show some results derived from a field survey using the policies. Chapter 5 describes some questionnaire-based measures implemented after the diagram that show the effectiveness of those measures.

## 2 Visualising Tool for Cooperative Relationship

This tool graphically represents answers to a questionnaire given by employees of an organisation regarding their cooperative relationships. We developed the questionnaire by combining an Interpersonal Solidarity Scale[7] and Bales' Interaction Process Analysis[8], and simplified it as much as possible. The questionnaire asks the employees to classify the other employees into any of the following five categories.

1. I don't know him or her.
2. I know his or her by sight.
3. I know his or her work and responsibilities.
4. I have talked about business with him or her.
5. I have worked with him or her.

The tool produces a matrix between employees which expresses their relationships.

**Static structure:** The tool renders employee's names as nodes and shows the cooperative relationships between employees as edges using a spring layout algorithm[9]. The spring strength is proportional to the value of the cooperative relationship. The edge is omnidirectional and the strength of an edge between the nodes is the average value of the relationships. A diagram shows that the cooperative employees are collocated and unknown employees are located far from each other. This tool can decorate nodes and edges in the following ways.

– The thickness of the edges can be proportional to the length. Users can easily see which relationships are strong in the diagram.
– Details of members of the group including their names and affiliation to the group are shown by colour in each node.
– The nodes can be clusterized based on edge betweenness[10] and the nodes in a cluster are painted with a colour. Users can identify isolated members.

**Dynamic structure:** The tool computes differences within the organisation and draws a comparison using two matrixes.

– It is possible to place two diagrams of cooperative relationships side by side, or switch over two drawings with a click.
– For each diagram, branches that show differences between cooperative relationships can be shown in red if their value is larger than the other branches, or in blue if their value is smaller than the other branches.
– A diagram showing only the difference between cooperative relationships can be displayed.

If nodes are divided between each group as shown in figure 1, it is expected that the distance between groups is large and there are no cooperative relationships between those groups. On the other hand, if employees of each group are uniformly spread as shown in figure 2, it is expected that there are cooperative relationships between the groups.

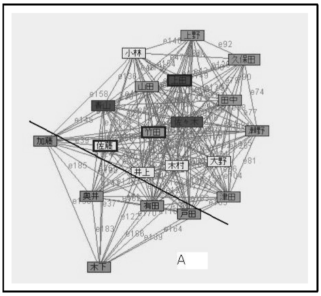


**Fig. 1.** Pre-poll cooperative relation

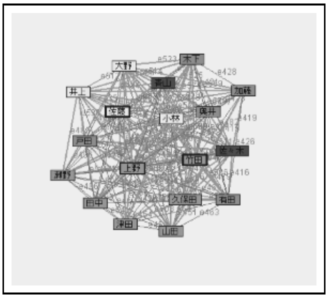


**Fig. 2.** Post poll cooperative relation

# 3 Survey Based on Field Research and Advance Questionnaires

## 3.1 Research Field and Research Methods

The organisation we surveyed implements software development projects based on various pieces of package software. It has about 200 employees in 7 groups, and we studied 4 of those groups. Of the employees, 23 people from 4 groups were interested in our research and we decided to do this research with them. The four groups deal with different work packages for their customers. Two or three people from one group worked together on a system development project, though usually they don't work together on projects and they don't talk or chat with employees of a different group in general.

We conducted two-hour interviews with 5 people from the 23. And we analysed them, derived measures to improve their communication, and helped them to implement those measures. We have continued to observe them. Independently from the interviews and observations, we gave questionnaires to all 23 employees and got answers from 21. We decided to use those 21 employees and their answers for our analysis and to make diagrams.

## 3.2 Cooperative Relations Based on Prior Surveys

Figure 1 shows result of the questionnaire was taken in prior surveys. Three teams B, C and D have a relatively uniform spread, but team A is located too far from the other three teams. We found a barrier between team A and the others. The average value was 2.7, and the average value with different team employees was 2.3. This means they know other team employees by sight but don't know what other team employees are doing in their work.

# 4 Field Survey and Analysis of Applied Measures

## 4.1 Analysis Model Overview

In this section we describe a mathematical model for transferring knowledge that we have developed with the purpose of having a KM mechanism or system take root in an organisation and obtaining guidelines to make it functional [2,3].

We have defined knowledge transfer by separating the parties into one which provides knowledge (the sender) and one which receives that knowledge (the receiver).

**Profit:** Direct benefit obtained by providing and receiving knowledge. The sender's profit includes incentives and the improvement of the sender's own skills. The receiver's profit includes improvement of work efficiency and improvement in probability of success as a result of receiving the knowledge.

**Cost:** Cost of providing and receiving knowledge. The sender's cost ($Cs$) includes the time required for creating documents, communicating verbally and working to provide information beneficial for the receiver. The receiver's cost ($Cr$) includes work to interpret or convert the received knowledge in order to use it.

**Barrier:** Something that influences knowledge transfer in relationships with others or in the environment. It can have either a positive or negative influence on the motivation to provide or receive knowledge. The barriers on the sender's side ($Bs$) include factors that affect the motivation to provide knowledge, such as the level of trust in the knowledge receiver and (competitive or cooperative) human relationships with the receiver. The barriers on the receiver's side ($Br$) include factors that affect the motivation to acquire knowledge, such as the level of trust in or preconceived notions about the knowledge provider as a result of past experiences of success or failure, the ability to search the system and the effort taken to listen to the knowledge provider.

Using these three factors, including the six parameters of $Ps$, $Cs$, $Bs$, $Pr$, $Cr$ and $Br$, we have modelled some aspects of knowledge transfer as follows:

- When knowledge is offered: $Ps - Cs > Bs$
- When knowledge is received: $Pr - Cr > Br$

The above expressions mean, if more profit than barrier remains after the cost is subtracted from the profit, knowledge transfer will occur.

### 4.2 Analysis of Field Research Based on Knowledge Transfer Model

From field research on employees, centring on interviews with them, we found that people who have problems contact those who seem to have knowledge when exchanging information with other employees. Further, when these survey results were applied to a mathematical model, the following conditions were found.

**Profit:** Employees don't think information and knowledge are worth providing to others because they don't get much reward. They get some profit if the knowledge solves the problems they are facing.

**Cost:** When in charge of dealing with the same customers and when solving issues related to those customers, explanations on those customers are not necessary. It is the same for general issues in which there is no dependence on customers or on packages, as software engineers exchange knowledge without much explanation. However, other work needs a lot of knowledge and a great many explanations are necessary.

**Barrier:** They don't trust each other much because they don't know each other very well. They show restraint because they don't know what each other is doing or each other's skill level.

When we apply this model to the organisation, we find that the organisation doesn't place much value on knowledge exchange and employees don't get much profit from providing knowledge. A lot of background knowledge is required to share knowledge of their work. And they don't know each other by sight. So we can say the barrier between employees is very large. We can suppose that there is insufficient communication for knowledge transfer and knowledge sharing.

### 4.3 Derivation of Measures

From the aforementioned situations, we suggested that the employees should have meetings for casual information exchange to develop relationships between employees in which they can help each other. In order to reduce barriers we tried the following variety of strategies to implement measures.

**System:** Following the instructions of managers, employees participate as part of the work. By positioning this work as work done during work hours, it becomes authorized work and the employees are not working as volunteers, and we thought this would reduce barriers in the system.

**Trust:** the purpose of this meeting is anxiety awareness where employees create relationships in which they can consult with each other, and we combined this meeting with self-introductions and icebreakers to build a relationship of trust.

**Sense of camaraderie:** We selected topics for the meeting very carefully, such as facilitation and reflection. These topics related to all employees because the topics are not related to their work and are what all employees require in general. All employees can empathise with each other and feel a sense of camaraderie. We took what was bothering them as the theme, and had them share their problems and also share a sense of camaraderie.

### 4.4 Results of Applying Measures

The information exchange meeting was held once a month for seven months. Figure 2 was made using the results of the questionnaire which we collected after this period. And also we got the following feedback from the employees.

- There were relationships in which people can consult with each other.
- It was good to know that employees in other groups had the same problems. a slight angle and to the right of you.
- We want to continue with implementation but we cannot do that yet by ourselves.

## 5 Validation of Findings from the Visualised Cooperative Relationships

### 5.1 Validation of the Visualised Relationships Before Implementing Measures

In the interviews, some employees said they know other employees only by sight. In figure 1, there is a gap between the upper right group and lower left group. This gap is backed by results from the interview. Also, the edges in figure 3 have values of more than 3.0. The gaps between the groups are shown in the diagram. The value of 3.0 means that they know what they are doing as their work. The diagram shows that they don't know about the other groups' work. The findings from this diagram are consistent with the fact that the employees said they required lots of background information to understand each other.

### 5.2 Validation of the Visualised Relationships After Implementing Measures

The number of samples was 21 and the average points rose from 2.7 (they knew each other by sight) to 3.1 (they know what each other are doing in terms of work). In the t-test, the value of $p$ was $0.000012 < 0.0001(0.01\%)$, which means the rise of the average value was statistically significant.

Each employee increased the value of his or her relationship with 6.8 employees on average. This means each employee built a deeper relationship with about 7 employees compared with before. The size of the circumcircle of employees in

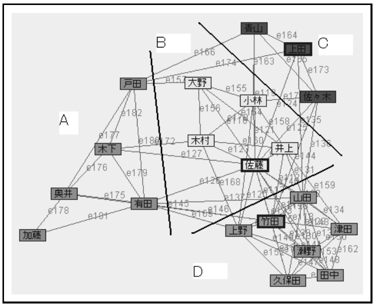


**Fig. 3.** The edges have values of more then 3.0

figure 2 is 17% smaller than the size of that in figure 1. This also backed the feedback from employees in section 4.4.

The purpose of the meeting was to develop relationships in which employees could consult with each other and we believe that purpose was achieved. This was also backed by the feedback from employees mentioned in section 4.4.

## 6 Summary

In this paper, we described a tool which gives a layout of employees as nodes and shows the cooperative relationships between them as edges. It visualises the static structure of cooperative relationships and the dynamic structure of relationships as a sequence of static structures and the difference of static structures. We also described how the tool works in a case study. The case study was based on our field research on an organisation that develops software. We compared the field research results with findings from the tool's visualisation based on a questionnaire given to the employees. The time needed to answer the questionnaire was between three and five minutes on average, and this is much shorter than the time taken for observations and interviews. But the many findings from the diagram drawn up using the results of the questionnaire back the findings of the survey results based on interviews.

As shown in section 4.4, in an analysis of the field research, we thought it was a big issue that employees knew each other only by sight. But such employees were in the minority and they did know each other, but they didn't know each other's work well. That is an example of how field research based on interviews can be misleading. The figure showed the barriers between the groups more notably. Therefore, managers in the field can use this tool and get suggestions on how to improve communication in their organisation. Looking at changes in diagrams is easier to understand than listening to opinions in the field through interviews, such as "I can now consult other team employees".

We will continue to investigate this organisation to help the measures and guidelines take root. In addition we would like to apply this tool to other organisations to enhance the features of this tool and improve its accuracy.

## References

1. Senge, P.M.: The Fifth Discipline: The Art & Practice of the Learning Organization. Currency (2006)
2. Ugai, T., Aoyama, K., Arima, J.: A mathematical model of knowledge transfer and case studies. The International Journal of Knowledge, Culture & Change Management 7, 9–17 (2007)
3. Aoyama, K., Ugai, T., Arima, J.: Design and evaluation a knowledge management system by using mathematical model of knowledge transfer. In: Apolloni, B., Howlett, R.J., Jain, L. (eds.) KES 2007, Part II. LNCS (LNAI), vol. 4693, pp. 1253–1260. Springer, Heidelberg (2007)
4. Wasserman, S., Stanley, K.: Social Networks Analysis: Methods and Applications. Cambridge University Press, Cambridge (1994)

5. Borgatti, S., Everett, M., Feeman, L.: UCINET IV Version 1.0 User's Guide. Analytic Technologies, Columbia, SC (1992)
6. Carley, K.M., Pattison, P., Breiger, R.L.: Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers. Natl. Academy Press (2004)
7. Rubin, R.B., Palmgreen, P., Sympher, H.E. (eds.): Communication Research Measures. Larence Erlbaum Associates, Mahwah (2004)
8. McGrath, J.E.: GROUPS: Interaction And Performance. Prentice-Hall, Englewood Cliffs (1984)
9. Eades, P.: A heuristic for graph drawing. Congressus Numerantium 42, 149–160 (1984)
10. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. Proc. Natl. Acad. Sci. 99, 8271–8276 (2002)

# Participatory Workshop as a Creativity Support System

Kosuke Numa, Kiyoko Toriumi, Katsuaki Tanaka,
Mina Akaishi, and Koichi Hori

The University of Tokyo, 4-6-1 Komaba, Meguro, Tokyo 153-8904, Japan
numa@ai.rcast.u-tokyo.ac.jp

**Abstract.** To support people's creativity, an information system is not the only way. The total design of the environment, the place and the tool for creation is required. In this research, we focus on *participatory workshops*, which are widely accepted for learning and creation. We propose a workshop as a creativity support system. The core elements of the workshop are participants, facilitator, information system, tasks, and place. Based on the model of *expression liquidization and crystallization*, we design the activities in a workshop. Collaborations in a workshop help to widen participants' views, and shared tasks and the place drives them to create new expressions. As a field trial of the proposed workshop, we designed and practiced *photo-attached acrostic workshop*. In this paper, we describe the designs of the new expression format called *photo-attached acrostics*, the workshop program, and the information system for supporting activities.

## 1 Introduction

When designing information systems, we need to examine how the system will be used by what kinds of users.

As for creativity support, many information systems to help people are developed recently [7,11], while many methods without information systems were developed [10,5] long before. This is not the problem like "which approach is better?" We need to choose and combine methods and systems according to the situation and the purpose. We regard a total environment including the place as *system*, not only an information system.

Not only in the engineering field, researches on creative activities are conducted. Especially in psychology and education, the style called *workshop* is now getting accepted [13]. By workshop here we mean a participatory and experiential group work-based style for learning and creation[1]. Learning and expressing will be activated in the group which shares a place and experiences.

It is often said that the knowledge productivity in a company is enhanced more efficiently by restructuring of organization — arranging the attitudes and the environment of people — than by installing intelligent systems. As for engineering

[1] Since the term *workshop* is used in several meanings, we use the phrase *participatory workshop* when emphasizing this characteristic.

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 823–830, 2008.

point of view, information system should be designed in the total environment where people use it. In this paper, we introduce our practice applied for people's art expressions.

Our targets, people's art expressions, include every artifact created or expressed by ordinary — non-professional — people. In many cases, the expressing processes are more important for ordinary people than the final expression, while the results are more important for professional artists. In this research, we aim to widen people's views to support expressing by a participatory workshop.

This paper is organized as follows: After describing related works in the next section, we provide our proposed framework. The design and results of our workshop are shown in Section 4 and we conclude the paper in Section 5.

## 2 Related Work

In the beginnings of 1990s, research area called creativity support was raised. In the area, problems like how computers can support human creative activity and what kind of creative activity can be supported were discussed.

Boden distinguished two sorts of creativity: H-creativity, which indicates historically new idea/concept formation, and P-creativity, psychologically new idea/concept formation in human minds [2]. In our research, we aim P-creativity support rather than H-creativity support. For ordinal people, our target users, what they express — externalization of internal nebulous thoughts — is more important than how they express — surficial originality of expressing techniques.

In psychology field, Guilford made the distinction between convergent and divergent thinking [3]. Our approach doesn't emphasize neither of them specially, but if daring to say, it matches divergent one. One of our aims is to support expressing, which seems to be a convergent process; but widening users' views and unsticking users' stuck thinking are more important.

## 3 Proposed Framework

### 3.1 Expression Liquidization and Crystallization

Expressions are interpreted based on the context — both the context inside the expressions and the context that the readers are placed in, and therefore the meanings or the values of the expressions depend on case by case. Cutting an expression off from a context and placing it into other context open new possibilities of interpretation. New interpretations stimulate people and cultivate new expressions. Expressions are placed into various context and again get other interpretation. In this section, we provide a model for this expression life cycle.

Some researchers modeled out people's knowledge and people's knowledge life cycle [8,12,9]. From the point of creativity support view, we have developed a cycle model which consists of the knowledge *liquidization* and *crystallization* processes [4]. Our basic idea is that knowledge is not a chunk of information, but emerges only within a certain context. We call the world in people's minds before

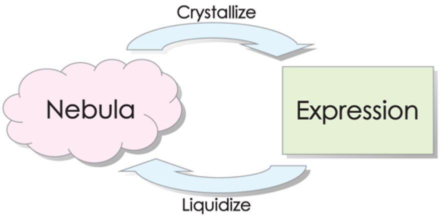


**Fig. 1.** Expression Liquidization and Crystallization

formed to a certain knowledge as *Nebulous World* or *Nebula*. In this research, we expand and apply the model to an expression life cycle. An expression is a special form of knowledge; it has one static form. But it is interpreted based on the context, which differs according to the situation or the state of the people and the expression. Contexts are relationships among units of partial expression and between them and units of external knowledge. These relationships always change. As *expression liquidization*, we call decomposition of expressions into units in proper granularity with every possible connection among each, and as *expression crystallization*, new expression formation from decomposed partial units based on new relationships within the context (Fig. 1).

When an expression is merged against an expression and when a context is merged against a context, the original context will be broken down and liquidization will be enhanced. Placing others expressions into a context changes the context and the values/meanings of expressions. Once a new expression created, it raises a new context, and then the new context stimulates her again.

### 3.2 Participatory Workshop

*Workshop* we focus on in this research is a participatory and experiential group work-based style for learning and creation. Workshops are held in various fields — arts, citizen-participatory town planning, and learning.

A workshop is arranged and organized by *facilitator*. The facilitator establishes tasks and prepares a place. Participants work together for the tasks in the place. Shared place and tasks enhance to form opinions and output expressions. In some case participants collaborate and in some case they compete.

Lave discussed the process of learning, creation, and consensus formation in a group called *Community of Practice* [14], where people share techniques, interests, or concerns. Commitment to the Community of Practice is activated by roles, which participants are required to play, such as a master and an apprentice [6]. This theory, *Legitimate Peripheral Participation*, explains participatory workshops gain participants' active commitments. A person, who plays a participant role, is requested to carry tasks out based on the program prepared by the facilitator.

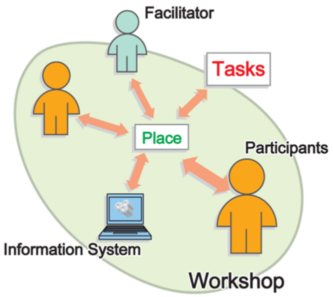


**Fig. 2.** Workshop as a System

### 3.3 Workshop as a System

Figure 2 illustrates the concept of *workshop as a system*. The core elements of a workshop are participants, facilitator, information system, tasks, and place.

Every person related to a workshop plays a role such as participant or facilitator. They gather in a place and work for given tasks. Information system supports people's activities in a workshop. These whole is a creativity support system called workshop; a participant herself is also a part of a system. A participant gets new ideas and creates expressions through tasks in a workshop, and at the same time, she contributes to others' creation as a part of the system.

The context where an expression placed in will be changed when one gets a new idea from an other participant or from an other's expression. Information system, however, can analyze and extract the structures of expressions based on the surface expressions, not on the people's subjective thoughts. Re-structuring expressions based on the changing context corresponds to liquidization. Balanced combination of outputs from information systems and interaction with other people will stimulate participants effectively. A workshop, at the same time, requests participants to output expressions as tasks in its program. The physical meeting promotes crystallization in this way.

## 4 Photo-Attached Acrostic Workshop

### 4.1 Workshop Design

We designed and organized a workshop as a field test. We prepared a place, established tasks, developed an information system, called for participation, and drived participants create expressions. In our first practice, we decided the theme as “Shonan” — the name of a region along a coast in central Japan, and called for participation to the people related to — e.g., living around, working around, or was born around — Shonan area. Through the workshop, participants discuss together and will get new opinions about the area.

In this research, we designed a new format of expression called *photo-attached acrostics* to highlight the process of decomposing and re-composing. Acrostic is

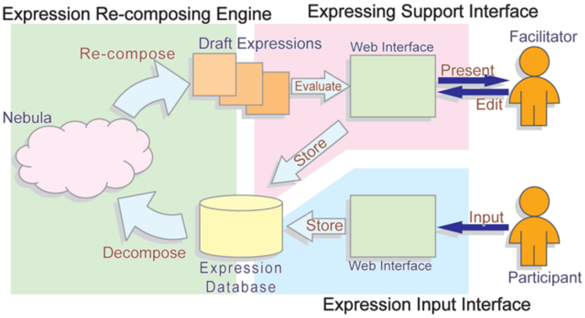


**Fig. 3.** Architecture of Information System

“a poem or other writing in an alphabetic script, in which the first letter, syllable or word of each line, paragraph or other recurring feature in the text spells out another message[2].” We modified it to include pictures for each sentence. Participants take and select photos, write sentences whose first letters match a message given. Here a pair of sentence and photo should correspond and both photos and sentences should be along a theme given.

In the workshop, participants create an acrostic using their own photos at first. Then next, they are divided into groups and collaborate to create new expressions by remixing their expressions. Collaboration with others will raise new context and stimulate participants. In the third step, they create expressions by themselves again, using all pictures used in the former steps. Patticipants are requested to place others’ (partial) expressions in their new expressions. We aim that participants form new opinions/ideas stimulated by others. At the same time, the workshop facilitator shows other new remixed acrostics using the developed information system described below. Here we aim to stimulate participants by (semi-)machinery generated — not manually created — expressions.

### 4.2 Installed Information System

The system consists of four parts (Fig. 3): expression database, expression input interface, expression re-composing engine, and expressing support interface.

The expression input interface is used in the former steps. Participants input their works, which are created in manual and analog manner. The expressing support interface shows the draft expressions, which are generated from the expression re-composing engine (see Fig. 4).

The expression re-composing processes are as follows:

- Decomposition phase
  1. Analyze the morphological structures of text.
  2. Calculate term relation weights and term weights.

[2] Acrostic – Wikipedia, the free encyclopedia: http://en.wikipedia.org/wiki/Acrostic

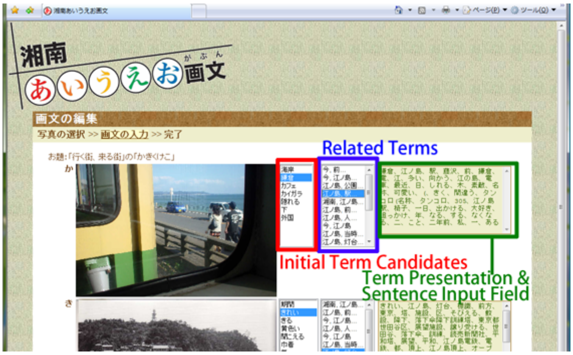

湘南あいうえお画文
かぶん
画文の編集
写真の選択 >> 画文の入力 >> 完了
お題:「行く街、来る街」の「かきくけこ」
Related Terms
Initial Term Candidates
Term Presentation & Sentence Input Field

**Fig. 4.** Screen Image of Photo-attached Acrostic Creation Support Interface

We use term dependency for term relation weights and term attractiveness for term weights [1]. Term dependency $td(t, t')$ from term $t$ to $t'$ is given by:

$$td(t, t') = \frac{sentences(t \cap t')}{sentences(t)} \tag{1}$$

Here $sentences(t)$ indicates the number of sentences in which term $t$ appears, and $sentences(t \cap t')$ is the number of sentences term $t$ and $t'$ appear at the same time.

Term attractiveness $attr(t)$ of term $t$ is a total of incoming term dependencies. $T$ is the set of all appearing terms.

$$attr(t) = \sum_{t' \in T | t' \neq t} td(t', t) \tag{2}$$

This directed network is a model of Nebula in this implementation.

- Re-composition phase
  1. Extract candidate terms according to their initial letters.
  2. Extract photos which include each term in 1.
  3. Evaluate photos.
     We define the weight $w_t(p)$ of a photo $p$ for term $t$ as follows:

$$w_t(p) = \sum_{t' \in T_p | t' \neq t} td(t, t') \cdot attr(t') \tag{3}$$

For each initial letter, the term candidates, their related terms, and attached photos are structured.

In the workshop, the facilitator shows semi-automatically generated expressions, which are editted in certain rules like choosing photos with the highest weights or the lowest. With these expressions, we aim to stimulate the participants by machinery generated context. Through this step, we observe the effects of the expression re-composing engine. After the workshop, the participants are asked to try the expressing support interface. We evaluate the interface from this test.

**Fig. 5.** Photo-attached Acrostic Workshop

### 4.3 Results and Discussions

The workshop was held at 8th and 16th December 2007 in Fujisawa city, the center of Shonan area, with nine participants. Most of their occupations are related to media activities or media literacy: information media-major students, an elementary school teacher, an art university professor, members of citizens' television at Shonan, and so on. While the youngest is an undergraduate student, a retired person is also included. Three are female, and six are male. The participants were divided into three groups and finally they made 30 photo-attached acrostics from 259 photos. Figure 5 shows the scenes in the workshop.

Through this workshop, we aim that participants exchange their knowledge and get new ideas through collaboration and competition. Most of the works from the latter steps were created by remixing others' former works. Several photos are used repeatedly by many participants. One participant, however, didn't change his mind finally. He preferred creating by himself rather than through collaboration. This fact shows our method is not almighty; this seems quite natural.

For the rest of participants, we found that collaborations in the shared place were effective. In the workshop, we prepared the tasks which consist of individual creations and collaborative creations. We expected that participants would change and expand their way of thinking through these tasks. From observation, it really worked, and more over, we found the participants frequently changed the balance of individual work and collaborative work even in one activity.

As for installed information system, we observed an interesting fact. One of semi-automatically generated expressions happened to have a similar story structure to a participant's one. We aimed to form a different context, but made a similar story. The user of the system, however, could create much more expressions in much less time. The outputs of the system are not always new, but the number of outputs can be large enough to stimulate the user.

It is important not to end a workshop alone, but to spread, repeat, and connect workshops. We paid a high cost for this preliminary workshop, but it is required to be held more easily so that many people can access workshops. A workshop is a closed system, but it can be opened by spreading an expression from the workshop to other workshops or by participating other workshops. When workshops are connected tightly and spread widely, our everyday lives will be covered with this creative environment.

## 5 Conclusion

In this paper, we claimed that information system should be designed together with the environment where it is used. We proposed a participatory workshop as a creativity support system. As a field trial of the proposed workshop, we designed and practiced *photo-attached acrostic workshop*.

## Acknowledgements

This work has been partially supported by a grant from the Japan Science & Technology Agency under CREST Project.

## References

1. Akaishi, M., Satoh, K., Tanaka, Y.: An associative information retrieval based on the dependency of term co-occurrence. In: Suzuki, E., Arikawa, S. (eds.) DS 2004. LNCS (LNAI), vol. 3245, pp. 195–206. Springer, Heidelberg (2004)
2. Boden, M.: The Creative Mind: Myths and Mechanisms. Basic Books (1991)
3. Guilford, J.P.: The nature of human intelligence. McGraw-Hill, New York (1967)
4. Hori, K., Nakakoji, K., Yamamoto, Y., Ostwald, J.: Organic perspectives of knowledge management: Knowledge evolution through a cycle of knowledge liquidization and crystallization. Journal of Universal Computer Science 10(3) (2004)
5. Kawakita, J.: The kj method: a scientific approach to problem solving. Technical report, Kawakita Research Institute, Tokyo (1975)
6. Lave, J., Wenger, E.: Situated learning: Legitimate peripheral participation. Cambridge University Press, Cambridge (1991)
7. Munemori, J., Nagasawa, Y.: Gungen: groupware for a new idea generation support system. Information and Software Technology 38(3), 213–220 (1996)
8. Nonaka, I., Takeuchi, H.: The Knowledge Creating Company. Oxford University Press, Oxford (1995)
9. Ohmukai, I., Takeda, H., Hamasaki, M., Numa, K., Adachi, S.: Metadata-driven personal knowledge publishing. In: Proceedings of 3rd International Semantic Web Conference 2004, pp. 591–604 (2004)
10. Osborn, A.F.: Applied Imagination: Principles and Procedures of Creative Problem-solving. Scribner (1957)
11. Ostwald, J., Hori, K., Nakakoji, K., Yamamoto, Y.: Organic perspectives of knowledge management. In: Proceedings of IKNOW 2003 Workshop on (Virtual) Communities of Practice within Modern Organizations, pp. 52–58 (2003)
12. Shneiderman, B.: Leonardo's Laptop: Human Needs and the New Computing Technologies. MIT Press, Cambridge (2002)
13. Stanfield, R.B.: The Workshop Book: From Individual Creativity to Group Action. New Society Publishers (2002)
14. Wenger, E., McDermott, R., Snyder, W.M.: Cultivating Communities of Practice. Harvartd Business School Press (2002)

# Readable Representations for Large-Scale Bipartite Graphs

Shuji Sato, Kazuo Misue, and Jiro Tanaka

Department of Computer Science, University of Tsukuba,
1-1-1 Tennoudai, Tsukuba, 305-8573, Japan
shuji@iplab.cs.tsukuba.ac.jp,{misue,jiro}@cs.tsukuba.ac.jp

**Abstract.** Bipartite graphs appear in various scenes in the real world, and visualizing these graphs helps improve our understanding of network structures. The amount of information that is available to us has increased dramatically in recent years, and it is therefore necessary to develop a drawing technique that corresponds to large-scale graphs. In this paper, we describe drawing methods to make large-scale bipartite graphs easy to read. We propose two techniques: "node contraction drawing," which involves collecting similar nodes and drawing them as one node, and "isosimilarity contour drawing," which puts clusters into an outlined area. We developed interactive user interfaces for the drawing methods and conducted an evaluation experiment to demonstrate the effectiveness of the proposed techniques.

**Keywords:** information visualization, graph drawing, bipartite graph, anchored map, clustering.

## 1 Introduction

With the spread of information services, the quantity of information that can be obtained has been increasing significantly. When people obtain knowledge from information, their understanding is strengthened by extracting and "visualizing" the information. Therefore, research on visualization techniques is being done in various fields. However, using common methods, the larger the amount of data becomes, the lower the readability is of the visualization results. Thus, we need to investigate new visualization methods.

The purpose of this research is to improve "readability" in the visualization of large-scale graphs. For the representation of a large-scale network using techniques used to draw anchored maps [1,2], we aim to improve readability by adding new representations of specialized techniques for large-scale graphs.

## 2 Previous Knowledge

### 2.1 Bipartite Graphs

A graph is a logical structure composed of a collection of nodes and a collection of edges that connect pairs of nodes, and is befitted to represent associations

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 831–838, 2008.

between some objects. A bipartite graph is a graph whose node set $V$ can be divided into two disjoint sets $V_1$ and $V_2$ such that every edge $e(\in E)$ connects a node in $V_1$ and one in $V_2$, and we have $G = (V_1 \cup V_2, E)$. As in the example of relations between customers and commodities and between academic papers and authors, bipartite graphs appear in various real-world scenes, and they also appear in research fields [3]. In this paper, we focus on the networks represented as bipartite graphs.

### 2.2 Anchored Maps

One of the most common methods used to draw graphs is the spring embedder model proposed by Eades [4]. This model calculates a stable state of a virtual physical model that has been constructed by regarding edges in a graph as springs.

An anchored map is an advanced form of the spring embedder model, and it restricts nodes in one of two node sets of a bipartite graph to certain positions.

- Nodes in $V_1$ are arranged at even intervals on the circumference, and then
- Nodes in $V_2$ are arranged by the spring embedder model, with a suitable position to represent their relationships to nodes in $V_1$

We call nodes in $V_1$ "anchors" and nodes in $V_2$ "free nodes." Edges are represented in a way that connects the anchors and the free nodes in a straight line. Fig. 1 shows the visualization result of the network of relations between academic papers and their authors by using anchored maps. Authors are represented by rectangles and are fixed as anchors.

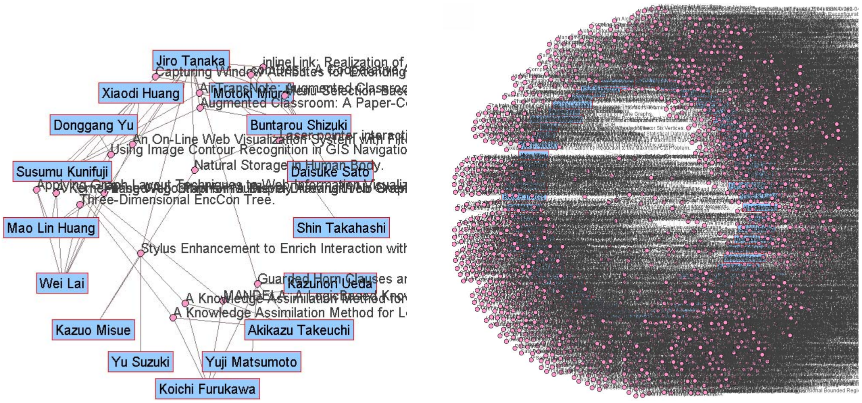
To improve the readability of anchored maps, anchors are arranged so as to reduce edge crossings and/or edge length[1]. However, when we draw a large-scale graph as an anchored map using the previous method [1], some problems occur in readability (see Fig. 1(b)). Large-scale graphs have a lot of nodes and edges for the drawing area, so the method of arranging the node layout makes it difficult to improve the readability. In previous studies a method of increasing the resolution of the display device and creating a very detailed drawing [5] was used, as well as a method of collecting several nodes and drawing them as one node [6,7]. However these methods do not allow the user to get a simultaneous overview as well as a detailed view of information. It is important for us to concisely draw a suitably sized graph that humans can recognize and understand.

## 3 Approach to Improve Readability

### 3.1 Readability of Graphs

When drawing a graph, the following situations should be distinguished: (1) we clearly see connections between nodes when there are few edge crossings,

[1] Refer to [1] for more information on how to determine the positions of anchors.


Jiro Tanaka
Xiaodi Huang
Donggang Yu
Susumu Kunifuji
Mao Lin Huang
Wei Lai
Kazuo Misue
Yu Suzuki
Koichi Furukawa
Yuji Matsumoto
Akikazu Takeuchi
Kazunori Ueda
Shin Takahashi
Daisuke Sato
Buntarou Shizuki
Motoki Miura
Augmented Classroom: A Paper-C
An On-Line Web Visualization System with Filt
Using Image Contour Recognition in GIS Navigatio
Natural Storage in Human Body.
Three-Dimensional EncCon Tree.
Stylus Enhancement to Enrich Interaction with
A Knowledge Assimilation Method for L

(a) anchors: 16, free nodes: 19

(b) anchors: 69, free nodes: 1891

**Fig. 1.** Example of drawing a co-authorship network using an anchored map

and (2) we are aware of some information when there are crowded nodes or a concentration of edges. The readability in conventional graph drawings was not sufficient to make those differences very clear and seems to have mainly focused on attributes like (1). We think that attributes such as those in (2) are also important to obtain knowledge from network information. Attributes like those in (1) and (2) are referred to as "legible" and "graspable" respectively to distinguish one from another in this paper. We consider them both to be important attributes for readability.

### 3.2 Clustering of Nodes

The proposed method first clusters the graph in order to analyze its structure. We tried to improve readability by enabling readers to change the visualized graph dynamically using the result of this clustering.

In most cases, clustering in graph drawing involves grouping nodes that logically have the same connected structure into a cluster. But if a graph becomes larger, the number of such nodes decreases, and thus, it is not effective to cluster them.

Clustering in this paper means to calculate the similarity of nodes from the connected edge relationships and to generate hierarchical clustering. We used the Jaccard index to calculate the degree of similarity between nodes and the nearest neighbor method as a clustering algorithm.

## 4 Node Contraction Drawing

We developed "node contraction drawing" so that the scale of the whole graph will be sufficient for readers to recognize the features of the clusters. With this method, pack nodes are expressed using the cluster information, with the purpose being to improve legibility.

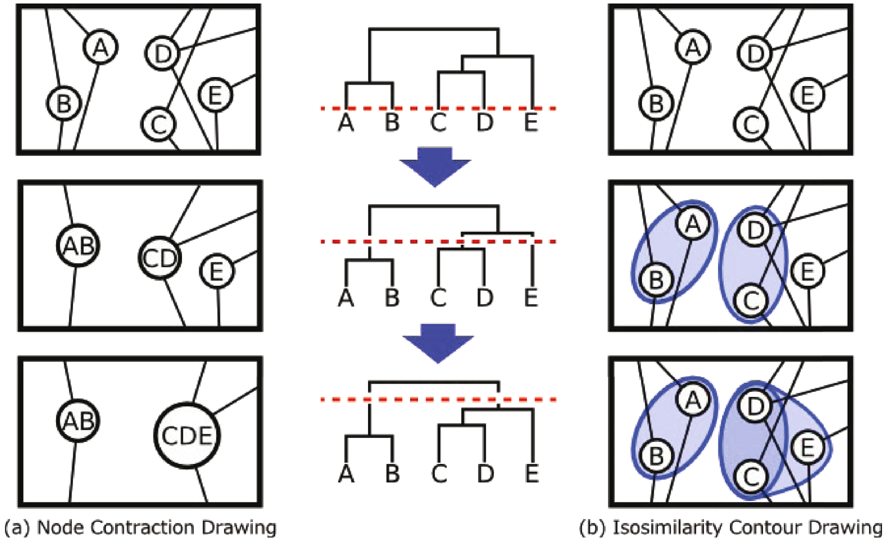

A
B
C
D
E
AB
CD
CDE
A B C D E
(a) Node Contraction Drawing
(b) Isosimilarity Contour Drawing

**Fig. 2.** Proposed representation model

Nodes are clustered according to the degree of similarity, from 0%–100%. A threshold value $t$ ($0 \leq t \leq 100[\%]$) is given, and our method draws clusters when nodes have a higher similarity value than $t$.

The middle column in Fig. 2 shows a clustering structure in the form of a dendrogram. The elements at the bottom of the dendrogram are nodes, and all the branch points correspond to clusters. The dotted line crossing horizontally through the dendrogram represents the threshold value $t$. In our method, showing clusters with a higher similarity than $t$ is synonymous with cutting the dendrogram horizontally and drawing only the node and the cluster connected to the edge of the cutting plane. Because the readers can change this $t$ and see the resulting visualized graph, we achieve the effect of dynamic clustering. This way, we can obtain the most suitable graph visualization for each reader.

A slider is employed on the interface that changes threshold value $t$; its maximum value is set to 100%, and the minimum is set to 0%. If the slider has a minimum value, only one cluster aggregating all nodes will be drawn.

An effective technique to show in detail the gaze point of clustered graphs involves displaying only the nodes that are focused on. One of the purposes of visualization is to obtain related information about the nodes that we pay attention to. In this case, the information about the nodes in the cluster is important. Also, it is possible to expand the cluster with the above-mentioned slider, but there is the possibility that unnecessary nodes may also be expanded. Because of that, we implemented a function that expands one cluster that is appointed at one time, so that the appointed clusters are expanded one by one.

As a result, the logical structure of the nodes being focused on can be seen, and the readers are assisted in their appropriate recognition.

## 5 Isosimilarity Contour Drawing

We wanted to develop a new method of improving readability that was different from dynamic modification of clustering level, and we came up with the idea of focusing on the reader's viewpoint. This method involves drawing a closed curve surrounding highly related nodes using the cluster information, and its purpose is to improve graspability. We named this method "isosimilarity contour drawing."

An effective way to show what humans fixate their eyes on is to indicate the important parts with a surrounding line. The formation of clusters can be recognized from the result even if clustering is not used. However, we aim to improve the recognition of the relations between nodes by actively displaying the result of clustering using a contour line.

The hierarchy clustering used in this research has a tree structure. It is possible to represent it using a nested structure. In an isosimilarity contour drawing, original nodes and edges are always drawn. If the dendrogram is cut horizontally, the clusters under the cut line are drawn. The cluster is represented as a closed curve that encloses all the included nodes.

The interface of this method also includes a slider, where we can interactively change the thereshold value to draw contours. When moving the slider in the negative direction, the number of contours increases until finally all nodes are surrounded by the biggest contours.

In this method, the contours are drawn as closed lines filled with a transparent blue color. By using a transparent color, overlapping of clusters can be represented. The rate of transparency is proportional to the degree of similarity of clusters. The higher the degree of similarity is, the lower the rate of transparency will be. Clusters that have closely related nodes and many layers are drawn with a denser color.

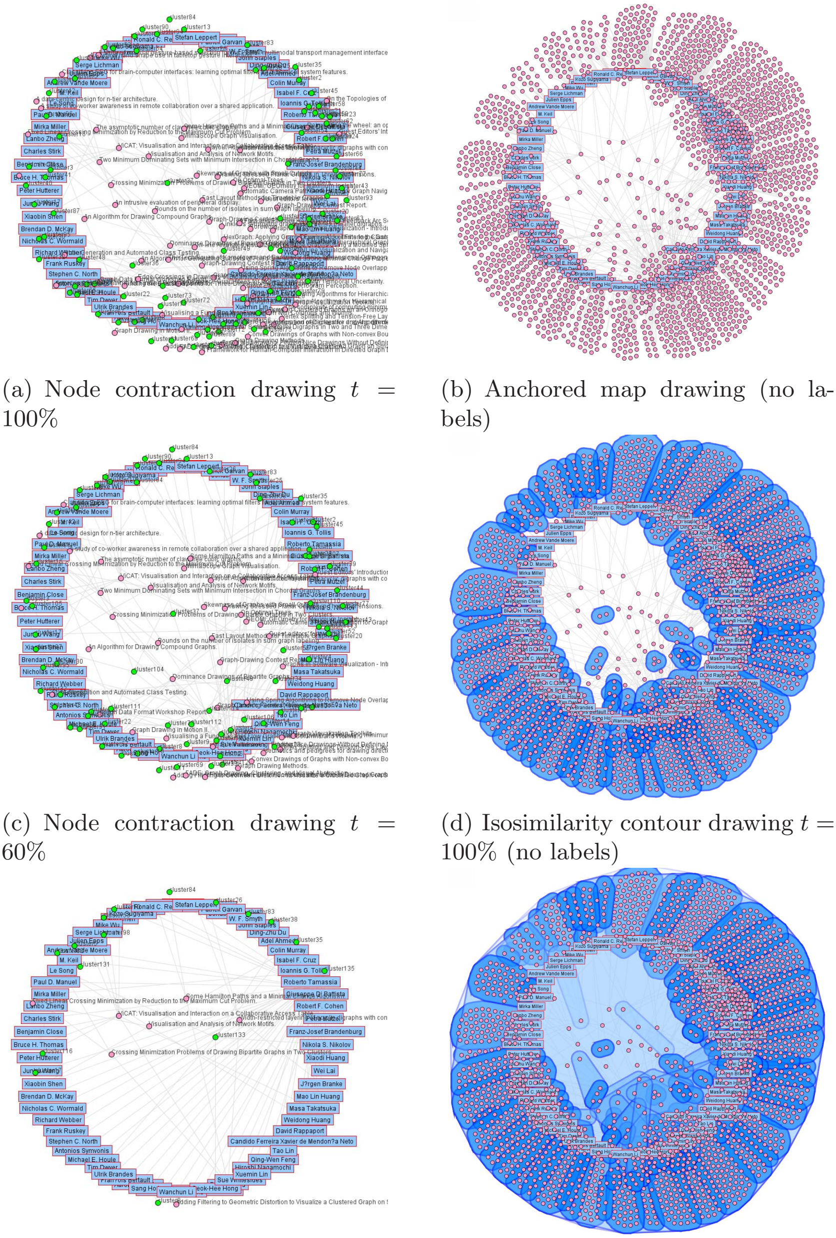
## 6 Drawing Examples

As an example, we visualized the network of coauthorships of papers obtained from DBLP[2] (anchors(authors): 69, free nodes (papers): 1891, edges: 2222). Because of the large number of nodes, when using a normal anchored map, it is difficult to understand the contents of the network (Fig. 1(b)).

The visualization results of the network using the node contraction drawing method are shown in Fig. 3(a),(c), and (e). When the node contraction drawing is activated and the slider is moved to 100%, the number of free nodes becomes about 150. The amount of free nodes having the same connections is very large, as papers tend to be written by the same group of coauthors. When the slider is moved to 50%, some papers remain in the center of the drawing area. These papers were written by an unusual group of coauthors, and we suppose that they connect several communities.

Fig. 3(d) and (f) shows the visualization result of the network using the isosimilarity contour drawing. The labels of free nodes have been omitted. The original

[2] http://dblp.uni-trier.de



(a) Node contraction drawing $t = 100\%$

(b) Anchored map drawing (no labels)

(c) Node contraction drawing $t = 60\%$

(d) Isosimilarity contour drawing $t = 100\%$ (no labels)

(e) Node contraction drawing $t = 50\%$

(f) Isosimilarity contour drawing $t = 0\%$ (no labels)

**Fig. 3.** Visualization results of relation between academic papers and their coauthors by using proposed representations

drawing as an anchored map without labels is shown in Fig. 3(b). We cannot see the contents of nodes from the figures, but we can recognize that the nodes with the highest similarity are placed at the circumference, nodes with higher similarity are at the bottom right in the center, and the nodes with lower similarity are in the center. We obtained the knowledge that this graph has a simple clustered structure.

## 7 Evaluation and Discussion

We conducted a user study to examine whether legibility and graspability were improved with our method. In the experiment, we used a graph of the relations between companies managing convenience stores and the number of their stores in prefectures in Japan, and we selected eight students (aged from 21 to 30) majoring in computer science as subjects. The subjects were asked to browse each graph drawn as a normal anchored map, the node contraction drawing, and the isosimilarity contour drawing. They were allowed to move the slider freely to change the value of $t$ so that their favorite drawing was provided. They were then asked some questions.

Table 1 and Table 2 present the results of the experiment. In the node contraction drawing, five people answered legible and their preferable threshold values are less than 100%. Based on these results, we believe that this method improved readability. The change of the threshold by the user was effective because the threshold value $t$ when the drawing is thought to be most legible is different for each subject. In the isosimilarity contour drawing, only two people answered legible, but five people answered that they obtained new knowledge. Therefore, we think that we were able to improve graspability with this method.

The node contraction drawing was designed to improve the legibility of the networks, but graspability might also be improved. A static image of a

**Table 1.** Experimental results of node contraction drawing

| evaluation item \ subject | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| Is the drawing legible? | **Y** | **Y** | **Y** | **Y** | **Y** | N | N | N |
| Preferable threshold value of $t$ | 80% | 70% | 85% | 74% | 75% | — | — | 70% |
| Did you obtain new knowledge? | **Y** | N | **Y** | **Y** | **Y** | **Y** | N | **Y** |

**Table 2.** Experimental results of isosimilarity drawing

| evaluation item \ subject | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| Is the drawing legible? | N | **Y** | ? | N | N | **Y** | N | N |
| Preferable thereshold value of $t$ | — | 90% | 94% | — | 100% | 80% | — | 85% |
| Did you obtain new knowledge? | N | **Y** | **Y** | **Y** | **Y** | **Y** | N | **Y** |

All questions were asked in Japanese. "**Y**" indicates a Yes answer, "N" is No, "?" means Not able to judge, and "—" means No answer. Preferable threshold value of $t$ means value of $t$ when drawing is thought to be legible.

contraction drawing with a fixed threshold value is not very effective. The process in which the users move the slider and change the visualized graph brings improved graspability.

The isosimilarity contour drawing was designed to improve the graspability of networks. Because this method is different from node contraction drawing, and original nodes remain without being combined, node relationships can be read from one fragmentary still image. We believe this method provides a good overview of the networks. This method is effective with a high threshold value and is suitable for observing the closely related nodes.

## 8 Conclusion

We proposed two visualization techniques to improve the readability of large-scale bipartite graphs and described their effectiveness. We developed a graph drawing tool by using hierarchical clustering. The tool provides a user interface to change the view of graphs from an overview to a detailed view and provides drawings with high readability. We developed the technique for anchored maps, but it is thought to be applicable to other styles of graph drawing, too. Some of our goals in the future include to find a way to avoid meaningless overlaps of contours and to make appropriate labels of nodes integrated by clustering.

## References

1. Misue, K.: Drawing bipartite graphs as anchored maps. In: Proceedings of Asia-Pacific Symposium on Information Visualization (APVIS 2006). CRIPT, vol. 60, pp. 169–177 (2006)
2. Misue, K.: Overview of Network Intormation by Using Anchored Maps. In: Apolloni, B., Howlett, R.J., Jain, L. (eds.) KES 2007, Part II. LNCS (LNAI), vol. 4693, pp. 1269–1276. Springer, Heidelberg (2007)
3. Newman, M.E.J.: Coauthorship networks and patterns of scientific collaboration. Proceedings of the National Academy of Sciences of the USA 101(suppl. 1), 5200–5205 (2004)
4. Eades, P.: A heuristic for graph drawing. Congressus Numeranitium 42, 149–160 (1984)
5. Lehmann, K.A., Kottler, S.: Visualizing Large and Clustered Networks. In: Kaufmann, M., Wagner, D. (eds.) GD 2006. LNCS, vol. 4372, pp. 240–251. Springer, Heidelberg (2007)
6. Frishman, Y., Tal, A.: Visualization of Mobile Object Environments. In: Proceedings of the 2005 ACM symposium on Software visualization, pp. 145–154 (2005)
7. Newbery, F.J.: Edge concentration: a method for clustering directed graphs. In: Proceedings of the 2nd International Workshop on Software configuration management, pp. 76–85 (1989)

# Evaluation of a Prototype of the Mimamori-care System for Persons with Dementia

Taro Sugihara[1], Kenichi Nakagawa[2], Tsutomu Fujinami[1], and Ryozo Takatsuka[1]

[1] Japan Advanced Institute of Science and Technology, Ishikawa, Japan
{sugihara, ryozo-t, fuji}@jaist.ac.jp
[2] Freelance programmer
macsi@m2.spacelan.ne.jp

**Abstract.** We aim to clarify the requirements for the mimamori-care system in the group home to support caregivers who take care of persons with dementia (PWD). We investigated the effects of the prototype, which consists of several cameras and monitors, by interviewing eleven caregivers. We found that the prototype system not only helps the caregivers to watch out the people with dementia but also removes some stress caused by taking excessive precaution to residents' behaviors. Caregivers however reported some concerns to the violation of privacy as for the caregivers themselves and the residents alike.

**Keywords:** Mimamori-care system, Caregivers' work stress, Persons with Dementia, and Group home.

## 1 Introduction

Japan is entering a super-aged society. There are approximately two million people with dementia (PWD). Group home is a carehouse for PWD, whose number has increased thirty-five times for these eight years, that is, from the year of 2000 to 2008. Despite the urgent need for care support systems and for eligible care workers in group homes, neither qualified persons nor any type of system has caught up the rapid social change.

We endeavor to develop a mimamori-care system to monitor the behaviors of the residents at group homes and to protect them from accidents. "Mimamori" is a Japanese expression, which means to watch someone or monitor something. "Mimamori-care" implies not only watching PWD but also supporting their autonomy to realize an adequate dementia care. A watch system for the elderly had been developed in the past but was not well accepted among PWD and caregivers due to privacy issues. Hence, it is necessary to solve these privacy issues and win their trust in order to induce them to use the digital equipment. Further, it is important to clarify the essence of dementia care. Hospitality and humanity constitute two pillars of caregiving. Therefore, the mimamori-care system should include these two components in the care of PWD as well as to assist caregivers. In order to solve the problems and to ensure reliable hospitality, we elaborate on a new concept; this system adopts the real world-oriented approach and is based on ubiquitous technology.

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 839–846, 2008.

Many researches and developments related to ubiquitous technology have been conducted. However, according to the valley of death [1] and the Darwinian Sea [2] metaphors, laboratory results cannot be implemented directly into the real world. There are methodologies such as human-centered design [3] and user-centered design [4] that reflect users' needs, improve usability, and facilitate modification in terms of custom-made design. However, dementia care is a difficult field. It depends heavily on contexts such as the person involved in, the type of care work, the equipment used, and the environment. The conventional approaches to dementia care have not taken them into account. Moreover, it is critical to understand the essence of dementia care with respect to factors such as the autonomy of PWD and their quality of life. Davenport mentioned the problem of infrequently-used systems about knowledge management; Knowledge applications have not been embedded into the flow of the job process with expert knowledge workers [6]. We defined this process as "the real world-oriented approach" and developed our ideas into the Mimamori-care system[5].

In this paper, we clarify the requirements for the mimamori-care system in the group home.

## 2 Mimamori-care System

Unless the technology and its users are mutually coordinated, the technology does not become an advantage for the users. We thus emphasize the viewpoints that the caregivers presented us with. That is, there should be a correspondence between the problem faced by the dementia caregivers and the functions implemented by the system.

PWD usually wander around group homes, which is a source of concern and anxiety for caregivers; the events may sometimes lead to a matter of life and death for PWD. Therefore, it is necessary to always know where PWD are and what they are doing. This system combines the functions of position detection and image capture by means of IC tags and cameras. Thus, caregivers can spot the locations and watch the behaviors of all the PWD through monitors. Figure 1 shows the schematic of the "mimamori-care" system. All the equipments are commercially available. We developed a program that allows the caregivers to communicate with each other. Originally, the intended functionality was to enable two or more caregivers to share a screen image, especially during the daytime. Each monitor may display different

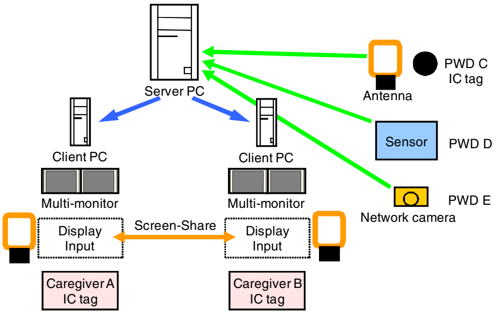


**Fig. 1.** Structure of the mimamori-care system

images dependent on the profile and context detectable based on the data from IC tags and sensors so that only the necessary items of information is provided. The Web context awareness technology [8] we had developed earlier was used for this objective. The server was developed in Visual Basic 6 using the API provided by the IC tag vendor. A Web browser was used to display the information to the client.

It is important that all the caregivers are notified of the collected items of information and that the priority of the work is clearly stated. Before introducing the system to a grouphome, our observation of the caregivers' activities revealed that their work involves a considerable amount of walking. To support their work, we introduced the prototype to two group homes. The prototype system is constructed from several cameras and several monitors to investigate issues of caregivers' work styles and

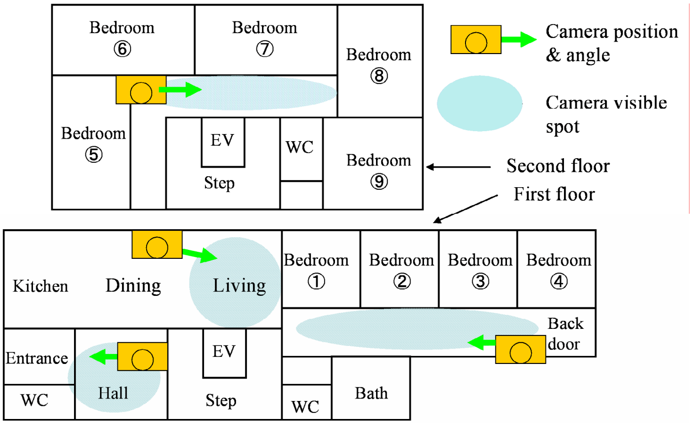


**Fig. 2.** The Arrangement of Cameras in Group Home A (GH-A)

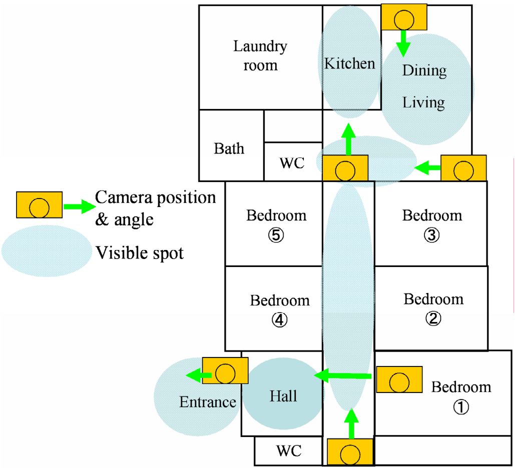


**Fig. 3.** The Arrangement of Cameras in Group Home B (GH-B)

stresses. To clarify the cause of changing work styles and stresses, we adopted only cameras and monitors whereas it is important for the concept of the "mimamori-care" system to include IC tags such as RF-ID. We embedded cameras in common spaces except the bathroom and restroom and monitors in the walls, in furniture, as well as in other different objects and places. We thereby created an ambient environment. We considered environments in which water is used, such as bathrooms, kitchens, and restrooms, and adopted waterproof monitors in such environments. We coordinated all these monitors with computers. The arrangements of cameras in the group homes are illustrated in Figure 2 and Figure 3. Group home A is a two-story house and Group home B is a single-story one.

Since the caregivers were inexperienced with computers, we designed touch-screen type monitors with a few buttons to enable even novice users to manipulate them appropriately.

## 3 Effects of the Prototype on Caregivers

### 3.1 Overview of the Research

To investigate the effects of the prototype, we interviewed eleven caregivers in the group homes. We believed it more effective to investigate the efficiency of the system through open-ended responses because the patient, caregiver, and situation change depending on the context [11]. We asked them which aspects they regard to be effective and what they think of the system. In particular, we paid attention to the difference in the uses of the system during daytime and nighttime caregiving. We also investigated how the prototype changed caregivers' mind after the introduction of the system.

Table 1 shows the overview of our interview. The group homes already reach at capacity, and total of caregiver is designated by law. In the daytime both GH-A and GH-B, caregivers work all over the home. They are busy with housekeeping, caregiving the residents who are PWDs, writing the records of the residents' behavior and so on. In the nighttime in both group homes, they mainly stand by around the living room. They may wander around and require an assistance for using the restroom.

The profile of the interviewees is illustrated in Table 2. We classified the caregivers' experience into three levels. Five caregivers are novice about the caregiving in the group home. Their experience is less than three years. Rest of the caregivers is classified as moderate level because they have worked in the group home with 3-7 years. Every worker has the qualification of caregiving and three caregivers in the GH-A have the nursing qualification.

### 3.2 Results and Discussions

Table 3 shows the evaluation points of the prototype and its typical answers. The two caregivers who have worked in both group homes mentioned a lot of good effects to the prototype. In the case of GH-A, they appreciate the effects of eliminating the blind spots, especially nighttime on the second floor. Before operating the system, caregivers had strongly worried when some of the residents is starting wandering on the

**Table 1.** Overview of this interview and group home

| | GH-A | GH-B |
|---|---|---|
| Residents | 9 | 6 |
| Total of caregivers | 9 | 5 |
| Intervewees | 6 | 5 |
| Caregivers in the daytime | 2 or 3 | 2 |
| Caregivers in the nighttime | 1 | 1 |
| Residential area | first and second floor | first floor |
| Start of operation | 2006.Dec. | 2005.Jan. |
| Video recording | none | on |
| Timing of interview | 2007.Jun. | |

**Table 2.** Profile of interviewees

| Intervewees | Experience level | Qualification of nurse |
|---|---|---|
| a1 | moderate | eligible |
| a2 | moderate | |
| a3 | moderate | eligible |
| a4 | low | |
| a5 | moderate | |
| a6 | low | eligible |
| b1 | moderate | |
| b2 | low | |
| b3 | moderate | |
| b4 | low | |
| b5 | low | |

second floor and how many times residents went to the restroom. When there was a trouble such as wandering, caregivers walked up from the living room to the second floor, which was a big blind spot, and helped or talked to residents. Caregivers could not feel relieved all night long and they could not concentrate on their works to be done. On the other hand, they have many works to do at the nighttime; For example, they have to write records about each residents' behavior (e.g. how many times residents went to the restroom at the daytime, how many times they took a walk, what kind of meal they ate), have to make a breakfast for residents of nine and themselves, have to help residents to change clothes, and have to clean rooms and corridor. They can now focus on the issues at hand.

Moreover, the prototype reduced their work stress by eliminating the blind spots, which brought their work style some change. Before operating the prototype, they felt the stress caused by undue interference. When residents attempted to go out somewhere, caregivers talked to residents too much than necessary. They were afraid of losing the sight of residents because they are not would like to wound by accidents. "We can do mimamori-work more effectively than before," one of the caregivers said after introducing the prototype. Recently, they felt that residents live more independent lives owing to the system. As a consequence, we have accomplished that our prototype provides caregivers with leeway to take adequate action to residents. It is important for caregivers to have the leeway. It reduces the number of errors in judgment, enhance the conversation time with residents, and improve the quality of life for

**Table 3.** Effects of the prototype

| | Evaluation point | Positive effects | Negative effects |
|---|---|---|---|
| GH-A | blind spots | - eliminated, especially at nighttime on the second floor. | none |
| | change in work style | - the caregivers easily monitor the activities of the PWD and take<br>- the caregivers keep the focus on the issues at hand, especially at nighttime<br>- going up and down stair has been reducing | none |
| | work stress | - the caregivers have attained the peace of mind because of eliminating<br>- the caregivers have obtained the relief because of reducing undue interference to the residents | none |
| GH-B | blind spots | - eliminated, especially at daytime on the hall and the entrance. | none |
| | change in work style | - the caregivers easily monitor the activities of the PWD and take<br>- the caregivers keep the focus on the issues at hand | none |
| | work stress | - the caregivers have attained the peace of mind because of eliminating<br>- the caregivers have obtained the relief because of reducing undue interference to the residents | - the video recording leads to high levels of stress to the caregivers<br>- the caregivers cannot rest in the break time because of the video recording |
| | video recording | - the caregivers can check where the resident are hit in the falling accident | - the caregivers are heavily stressed because of violation of the privacy rights with themselves, co-workers and residents. |

residents. If caregivers are not provided with a leeway, they will lose not only motivation but also creativity. In the case of GH-A, caregivers do not mention about bad effects.

Although almost good effects for caregivers in the GH-B are as same as GH-A, the prototype works more effectively in the daytime. One of the residents has a habit of taking a walk, and GH-B do not lock the door to the entrance. Before operating the system, caregivers sometimes passed over resident's outgo. It is difficult to notice the behavior because caregivers may be at a location from which they cannot observe the residents. After installing the system, the caregivers can monitor the activities of the PWD and take adequate reaction. They said "the system is half worth of a caregiver" and "my sight is enhanced." However, they also reported of negative effects of video recording, which violates the privacy rights of themselves and residents. Despite the fact that they recognize the advantage of video recording (e.g. the caregivers can check where the resident is hit in the falling accident), they are heavily stressed (e.g. the caregivers cannot rest in the break time). Earlier, some caregivers experienced stress while caregiving residents. However, this system decreases the guilt for residents' privacy. The problem is very important because it becomes inhibition effect to the caregivers' creativity. If caregivers strongly feel that they are under constant surveillance by someone, they would feel threaten. There is a trade-off between this problem and the advantage of the video recording. Therefore, it is necessary to pay attention to caregivers' feelings continuously.

## 4 Related Works

In Japan, one of the most well-known researches using radio frequency identification (RFID) is a tracking system for elementary school pupils [9]. These systems typically use active RFID by which pupils can be automatically identified at the school gate. On the other hand, researchers also used Global Positioning System (GPS) to verify the location of wandering the elderly [10]. The problem, however, occurred when a resident removed the device due to its stumbling block. In the mimamori-care system, we used a 13.56 MHz passive RFID tag, which can be used permanently and does not require batteries. The RFID we used effectively communicates within the range of 70 cm. As our target is a typical house, this range does not pose a problem. By installing antennas in passages and on doors, we can effectively use this RFID. Using the above-mentioned RFID, we experimented whether movements could be detected in each room.

For the nursing action in hospitals, the e-nightingale project, which focused on recording with wearable sensors, was proposed [11]. Our target is not to record the nursing action in hospitals but to support PWDs' self-reliant life in group home.

## 5 Conclusions

This paper presents the mimamori-care system and its effects on caregivers. Our results show that the system was effectively used in both daytime as well as nighttime caregiving at two group homes. Despite the system makes peace of mind, which is one of the most important factors for the mimamori-care works, the system caused some stress to caregivers, who are concerned with violation of the privacy rights. One of the purposes of caregiving is the independence of PWD; this system encourages PWD to be more independent. Although this system does not intend to be a substitute for caregivers, it provides them with considerable support. The reason for the positive vibes in group homes is the comfortable, homelike environment. In these surroundings, PWD are not only caregiven but are also able to engage in other activities of daily life. Caregivers are enabled to peacefully make improvements in the quality of care.

**Acknowledgments.** Our research is partly supported by the fund from Ministry of Education, Culture, Sports, Science and Technology, Japan, under the name of Cluster for Promotion of Science and Technology in Regional Areas.

## References

[1] Wessener, C.: Public/Private Partnerships for Innovation, US National Academy of Sciences. In: OECD Workshop (2001)

[2] Branscomb, L.M., Auerswald, P.E.: Between Invention and Innovation An Analysis of the Funding for Early Stage Technology Development, Report to the Advanced Technology Program, NIST, US Department of Commerce, GCR-02-841 (2002)

[3] Takatsuka, R., Fujinami, T.: Aware Group Home: Person-Centered Care as Creative Problem Solving. In: Khosla, R., Howlett, R.J., Jain, L.C. (eds.) KES 2005. LNCS (LNAI), vol. 3684, pp. 451–457. Springer, Heidelberg (2005)

[4] ISO13407, Human-centred design processes for interactive systems (1999)

[5] Nakagawa, K., Sugihara, T., Koshiba, H., Takatsuka, R., Kato, N., Kunifuji, S.: Development of a Mimamori-Care System for Persons with Dementia Based on the Real World-Oriented Approach. In: Apolloni, B., Howlett, R.J., Jain, L. (eds.) KES 2007, Part II. LNCS (LNAI), vol. 4693, pp. 1261–1268. Springer, Heidelberg (2007)

[6] Davenport, T.H.: Thinking for a Living: How to Get Better Performances and Results from Knowledge Workers. Harvard Business School Press (2005)

[7] Kanai, H., Nakada, H., Turuma, G., Kunifuji, S.: An Aware-Environment Enhanced Group Home. In: International Workshop on Smart Home (IWSH 2006) in conjunction with International Conference on Hybrid Information Technology (2006)

[8] Nakagawa, K., Kato, N., Ueda, Y., Kunifuji, S.: A proposal and an implementation of a context-awareness system for web-based collaboration. IPSJ 47(7), 2081–2090 (2006)

[9] Yamada, I., Shiotsu, S., Itasaki, A., Inano, S., Yasaki, K., Takenaka, M.: Secure Active RFID Tag System. In: Proceeding of UbiComp 2005 Workshop on UbiComp Privacy "PRIVACY IN CONTEXT" (2005)

[10] Shimizu, K., Kawamura, K., Yamamoto, K.: Location System for Dementia Wandering. In: Proceedings of the 22nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 4928-92332.pdf(CD-ROM) (2000)

[11] Kuwahara, N., Noma, H., Kogure, K., Hagita, N., Tetsutani, N., Iseki, H.: Wearable Auto-Event-Recording of Medical Nursing. In: Proceedings of INTERACT 2003 (2003)

# Aware Group Home Enhanced by RFID Technology

Motoki Miura[1], Sadanori Ito[2], Ryozo Takatsuka[1], and Susumu Kunifuji[1]

[1] School of Knowledge Science,
Japan Advanced Institute of Science and Technology
miuramo@jaist.ac.jp, ryozo-t@jaist.ac.jp, kuni@jaist.ac.jp
[2] Course of Ubiquitous and Universal Information Environment,
Department of Computer and Information Sciences,
Graduate School of Engineering, Tokyo University of Agriculture and Technology
sito@cc.tuat.ac.jp

**Abstract.** In Japan, the number of *group homes* offering home-like care for elderly persons suffering from dementia has increased considerably. Even though the lesser number of people residing in a group home is suitable for family-like care, the shortage of caregivers increases the burden, especially during the night. To supplement this lack of attention, we developed floor mats embedded with RFID antennae and slippers with RFID tags. These can help caregivers be aware of the activities of persons suffering from dementia, by specifying whether an individual has passed over a sheet in a particular corridor. This not only helps the caregivers understand such persons by reviewing their activities, but also keeps them informed about the person's current activities.

**Keywords:** Ubiquitous home, Location awareness, Elderly persons care.

## 1 Introduction

A group home is a type of home-like care facility in Japan for elderly persons suffering from dementia. Since Japan is one of the fastest aging societies, developing such facilities where elderly people can live safely and at ease is crucial. In an aging society, people's need for care services are greater but the number of large-scale care facilities are still limited. Consequently the demand for group homes for the elderly has increased.

A group home is a relatively small facility based on the principle of normalization; it offers family-like care to enrich the quality of life by taking into account the personalities of individual elderly persons. A group home is suitable for implementing this philosophy because the number of persons residing there is relatively low. In the daytime, up to nine elderly persons are cared for by three caregivers. During the night, however, only one caregiver provides support for all nine individuals. It is therefore difficult for the caregiver to deal with concurrent events, and to memorize all behavioral events.

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 847–854, 2008.

To relieve the burden of the caregiver, our research group is developing “Aware Group Home.” Although, its concept is similar to that of Aware Home Research Initiative [1], we focus on providing support for caregivers living with persons suffering from dementia [2]. We built an experimental group home named AwareRium with built-in sensors such as ultrasonic position detectors, active RFID readers, and floor-pressure sensors. Although these advanced sensors are effective in augmenting the functionality of group homes, it is difficult to install them in real group homes. Moreover, although the floor-pressure sensors can detect events such as the approach of a person to a particular area, they cannot distinguish between individuals, unless identification tags are attached to them.

To solve these problems, we developed a simple radio frequency identification (RFID) mat sensor system. Our system detects a person’s approach to a specified area, and stores the information for future reference or to provide information directly to the caregiver. In this paper, we describe the system and lessons learned from our experimental results.

## 2 RFID Mat Reader and Its Applications

Here, we describe the requirements, implementation, applications, and installation of the system for a group home.

### 2.1 System Design Requirements

Improving care service based on person-centered concepts requires taking individuals into consideration. Since our system supports caregivers who provide person-centered care, it must identify individuals, and sensors that simply detect people passing or moving about are insufficient. Attaching personal identification devices to the elderly person’s body, however, should be avoided. Battery-powered devices increase the work of the caregivers-they must change or charge the batteries regularly. In addition, we cannot fully modify the existing group homes, because many group homes are established after minor renovation of conventional houses. After considering these requirements and our own observations, we chose RFID technology to detect personal activities of elderly persons in group homes by embedding tags in their slippers. The tag is enough small to be embedded, and elderly persons always wear slippers in the home. The passive RFID tags do not require additional work, such as charging batteries, thus minimizing additional maintenance tasks for caregivers.

### 2.2 Usage Situations

Caregivers usually provide assistance to the elderly after considering their individual abilities and personalities. Understanding personal activities based on location-aware services is advantageous to the caregivers due to the following points:

1. At night, some elderly persons frequently visit the restroom. Relevant notification along with identification helps the caregivers determine whether the

individual requires assistance. In addition, the caregivers will know when the individual would normally return from the restroom. Identification can also be used to record the frequency of restroom visits for each person. In case of higher frequency, the caregiver should be notified about the patient's shortage of sleep.

2. Sometimes, people with dementia perform redundant activities such as brushing their teeth every 10 minutes. If the system generates a person's activity log, patterns could be identified that help in determining the person's basic state of mind and devising countermeasures.
3. An irregular action such as roaming around or walking down a corridor at high speed can represent a state of high emotion. Our system can detect these irregular actions and inform the caregivers

In addition, the caregivers have to maintain "care notes" to record activities. The activity log generated by the system can provide fundamental and objective data for the note records.

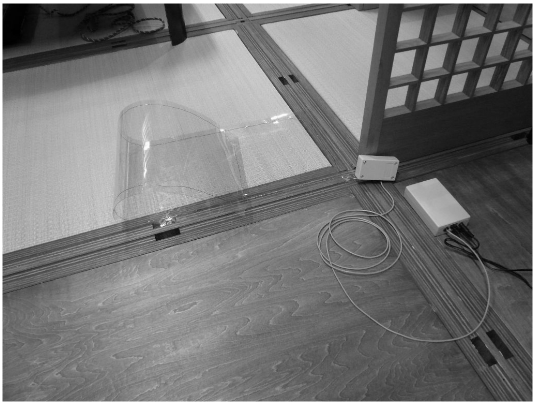


**Fig. 1.** RFID antenna sheet, regulation circuit box, and reader box

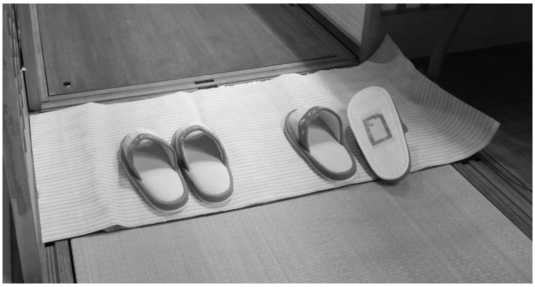


**Fig. 2.** Slippers with RFID tags (in actual operation, the tags are embedded)

### 2.3 Implementation

The details of our initial prototype of the RFID mat made using copper tape are described in [3]. To reduce the effort of preparing multiple antenna sheets, we employed enamel wire stuck on a thin plastic film (Fig. 1). We prepared three types of antenna sheets of different sizes (300mm × 400mm, 300mm × 650mm, 300mm × 1200mm) to suit different locations. The differences in operation due to antenna size are regulated by variable resistances in the regulating circuit box (90mm × 50mm × 26mm). The RFID reader box (125mm × 80mm × 32mm) consists of an HHPA module developed by Sobal Corp.[1], a power supply circuit, and a LAN module (XPort[2]). The HHPA module is compatible with ISO15693-2 tags, and can provide an output of 1W. An RFID tag sheet (OMRON V720SD13P01) was embedded into the slippers (Fig. 2). The reader can detect slippers within a range of 5 to 10 cm. Since the stride length of an elderly person is shorter than that of a younger person, the performance of the reader/antenna is sufficient for group homes.

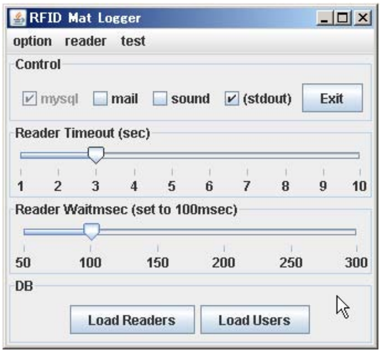
We developed a simple event logger that monitors the detection events of the RFID tags in Java (see Fig. 3). When an RFID tag is detected by an RFID reader, a data pair (tag ID, reader ID) is generated and registered as an "enter" event. When the tag leaves the field, the event "exit" is registered. To prevent irrelevant detections or missing events, the user can set "Reader timeout" with the slider. When the reader detects an "enter" event within the timeout, the last "exit" event is canceled. Thus, the event logger stores the data (reader ID, tag ID, date/time entered, date/time exited) in a MySQL table. The "Reader waitmsec" slider allows the user to set a polling interval time for each reader. The "Load Readers" and "Load Users" buttons restore the descriptions of readers and tag IDs from the database, respectively. The application also has functions to add extra readers dynamically, which is beneficial in maintaining readers with less effort.

Fig. 4 shows a simple data browser for the log. The browser loads the log from the MySQL table and places the person icons on a floor map of the group home. The caregiver can review the past activities of the persons by moving the knob on the slider located below the window. Thus, caregivers can review past events, even though they were engaged in caring for another person.

Since the caregiver is usually engaged in some work, it is difficult for him/her to watch a status window for each event. Therefore we implemented a voice announcement function. When a person needing assistance is detected in the corridor, the system announces: "Mr. XX (owner of the slippers) is in corridor (location of the antenna mat)." This effectively allows caregivers to know the current situation without looking at the monitor, especially in case 1 described in section 2.2. The announcement function is customizable for a real-word situation. For example, it can be programmed to report the frequency of using the restroom. The voice announcement function was developed using Galatea Toolkit [4].

[1] http://www.sobal.co.jp/rfid/rfid-hhpa.html

[2] http://www.lantronix.jp/products/ds_xport.shtml



☑ mysql □ mail □ sound ☑ (stdout)

**Fig. 3.** Event logger application

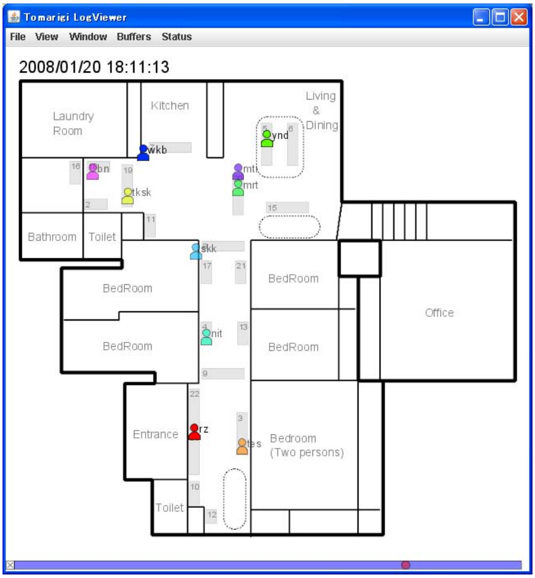


**Fig. 4.** Log browsing interface

We also prepared a web interface for caregivers to observe the transitions of personal activities (Fig. 5). Using this view, the caregivers can be aware of the amount and trends of the activity of each person.

## 2.4 Installation and Operation

To validate the effectiveness of our system, we deployed the antennas and readers in an actual group home, named Tomarigi. Six elderly persons with dementia live in the home. Even though the RFID mat sheet is thin (thickness, 2 mm),

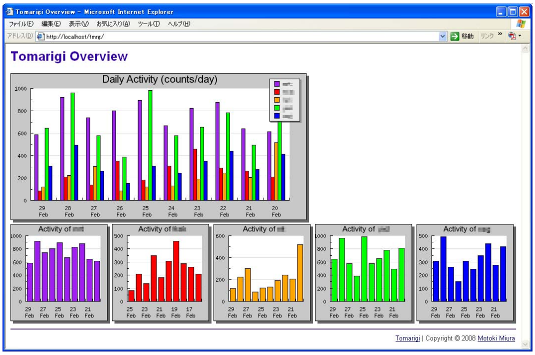

Tomarigi Overview – Microsoft Internet Explorer
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
アドレス(D) http://localhost/tmrg/
移動 リンク
Tomarigi Overview
Daily Activity (counts/day)
1000
800
600
400
200
0
29 Feb 28 Feb 27 Feb 26 Feb 25 Feb 24 Feb 23 Feb 22 Feb 21 Feb 20 Feb
Activity of [redacted]
Activity of [redacted]
Activity of [redacted]
Activity of [redacted]
Activity of [redacted]
29 Feb 27 Feb 25 Feb 23 Feb 21 Feb
25 Feb 23 Feb 21 Feb 19 Feb 17 Feb
Tomarigi | Copyright © 2008 Motoki Miura

**Fig. 5.** Activity transition graphs

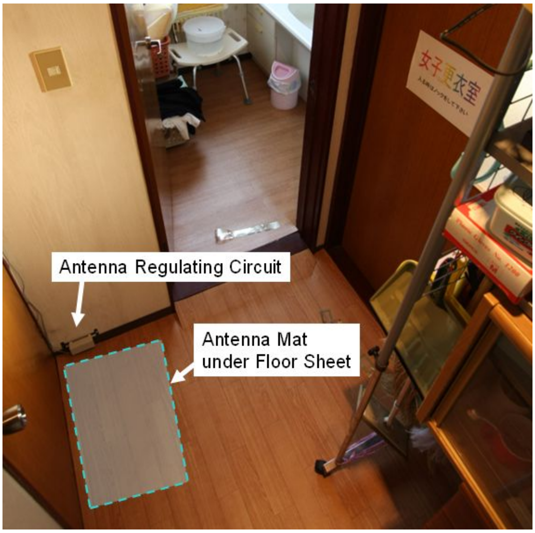


**Fig. 6.** Deployment of an antenna in front of a restroom (No. 2 antenna in Fig. 4)

placing the sheet directly on the floor of the common space could cause patients to trip. Therefore, we placed the antenna sheet under a carpet (Fig. 6), which covers the whole floor of the room and the corridor. Fig. 7 shows the corridor and highlights the antenna positions, and Fig. 4 illustrates the positions of antennas in the group home. Note that the dotted shapes in the figure represent a dining table and two sofas. The boxes containing the regulation circuits are attached to the wall near the mat antenna. The RFID reader boxes are also attached to the wall, below the ceiling. We connected each reader box and hub with network cables. We did not adopt wireless LAN since we found that a microwave oven disturbs the network. One PC collects all data from the readers by polling every

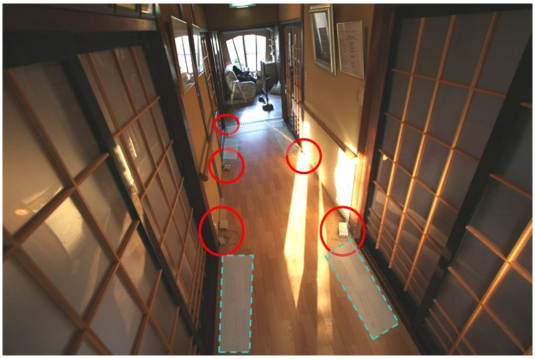


**Fig. 7.** Deployment of an antenna in a corridor (before mounting RFID reader boxes)

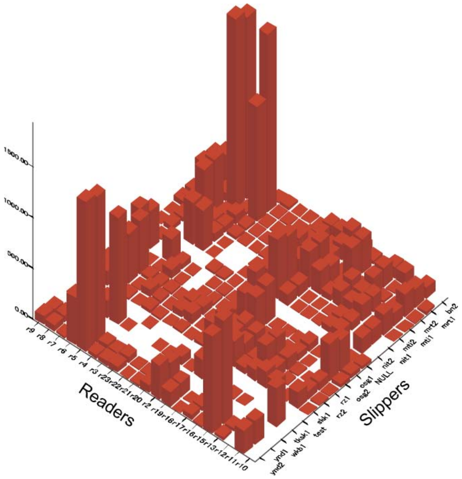


**Fig. 8.** Histogram of detected numbers for each reader/slipper

100 msec. The log browsing interface and voice announcement function were not implemented as part of the experiment.

We made tag-embedded slippers for both the elderly persons and caregivers. Seven caregivers and five elderly persons, not including one person who used a wheelchair, wore the slippers. To prevent mix-ups, the slippers were of different colors. We began logging data from January 2008. We can collect 4000 event logs per day. The slippers were seldom misplaced since each person recognized them as personal belongings. We have not analyzed in terms of long-term statistics, although a histogram (Fig. 8) reveals a tendency of staying near the dining table (readers 5 and 6) or the sofa (reader 15) for a long time. Thus, we confirmed that our approach for detecting behavior with RFID mat sensors is applicable to group homes.

## 3 Conclusion and Future Work

We described the design of an enhanced group home with RFID mat antennas to relieve the burden of caregivers. We described the system, which uses the activity data of elderly persons, and its application, and also confirmed the feasibility of our approach.

In future, we plan to continue collecting data and providing the services for caregivers mentioned in this paper. In particular, we want to identify gradual tendencies, which are difficult for caregivers to recognize. We will reveal the effectiveness of the logging system and the applications based on the caregivers' comments. The activity logs can be used not only for improving daily care but also for training of novice caregivers.

## Acknowledgment

The RFID technologies are presented by Sobal Corporation. Our research is partly supported by a fund from the Ministry of Education, Culture, Sports, Science, and Technology of Japan, under the name of "Cluster for Promotion of Science and Technology in Regional Areas."

## References

1. Kidd, C.D., Orr, R.J., Abowd, G.D., Atkeson, C.G., Essa, I.A., MacIntyre, B., Mynatt, E., Starner, T.E., Newstetter, W.: The Aware Home: A Living Laboratory for Ubiquitous Computing Research. In: Streitz, N.A., Hartkopf, V. (eds.) CoBuild 1999. LNCS, vol. 1670. Springer, Heidelberg (1999)
2. Kanai, H., Nakada, T., Turuma, G., Kunifuji, S.: An aware-environment enhanced group home: AwareRium. In: International Workshop on Smart Home (IWSH 2006) in conjunction with International Conference on Hybrid Information Technology (2006)
3. Miura, M., Ito, S., Kunifuji, S.: Development of RFID Mat Sensor System for Person-Centered Care in Group Home. In: Proceedings of the 2nd International Conference on Knowledge, Information and Creativity Support Systems(KICSS 2007), November 2007, pp. 59–62 (2007)
4. Kawamoto, S., Shimodaira, H., Nitta, T., Nishimoto, T., Nakamura, S., Itou, K., Morishima, S., Yotsukura, T., Kai, A., Lee, A., Yamashita, Y., Kobayashi, T., Tokuda, K., Hirose, K., Minematsu, N., Yamada, A., Den, Y., Utsuro, T., Sagayama, S.: Open-source software for developing anthropomorphic spoken dialog agent. In: Proceedings of PRICAI 2002, International Workshop on Lifelike Animated Agents, August 2002, pp. 64–69 (2002), http://hil.t.u-tokyo.ac.jp/~galatea/

# A Tabletop Interface Using Controllable Transparency Glass for Collaborative Card-Based Creative Activity

Motoki Miura and Susumu Kunifuji

School of Knowledge Science,
Japan Advanced Institute of Science and Technology
miuramo@jaist.ac.jp, kuni@jaist.ac.jp

**Abstract.** Conventional tabletop systems have focused on communication with virtual data, using *phicons* or physical objects as handles. This approach is versatile, given the full use of a horizontal display. However, we consider that an another approach can be formulated that can support normal specific tasks on a table. We have developed a card-handling activity environment enhanced by a tabletop interface. We use a glass tabletop with controllable transparency to improve surface scanning and the display of supplemental data. We describe the architecture of the tabletop system and its design criteria. Due to its simple configuration, this tabletop system can handle a large number of paper cards, as used in the KJ method. Therefore, our system can be used to enhance card-based tasks by showing additional data, and it provides the ability to review transactions by recording the tasks.

**Keywords:** Knowledge Creation, Physical Label Works, Ubiquitous, Paper and Pen.

## 1 Introduction

A great deal of research has been conducted on tabletop interfaces. In general, the term *tabletop* represents a table installed as a relatively large horizontal display, which is intended to facilitate collaboration among multiple users surrounding the table. The characteristics of the tabletop interface are: (1) multiple users can join the workspace on equal terms, and (2) the users can place objects on the tabletop and interact with the physical objects as well as virtual objects shown on the surface. Since these characteristics are promising for improving collaboration, considerable research is being conducted on the tabletop interface.

The philosophy of the tangible user interface (TUI) and the concept of extending table-based activity into the virtual world are promising. These concepts and technologies can be gradually and naturally applied to extending the conventional tasks performed on a table, using ubiquitous computing. We chose a knowledge creation activity that involves handling a set of paper cards, in the manner of the KJ method, and developed a tabletop system called AwareTable to gradually extend the activity.

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 855–862, 2008.

The concept of “gradual extending” in this paper refers to a phenomenon in which the system provides intermittent support, when necessary. In the case that we can accept the intermittent system support, the system can be realized by simple configuration. AwareTable provides two time-sharing functions: a scanning tabletop and displaying of additional data. We aim to augment paper card-handling activity without attaching any special devices to either the cards or the user’s hands.

## 2 AwareTable

Here, we describe the functions, design guidelines, and implementation of AwareTable.

### 2.1 Functions and Design Guidelines

The primary function of AwareTable is to record paper card transactions during card-handling activities for future reference and review. To fulfill this purpose, the system must be able to detect individual paper cards, and record the location of the cards on the table.

DigitalDesk[1,2] and EnhancedDesk [3] capture photo images of the tabletop with a camera located over the table to recognize objects and fingers. This approach is straightforward, but the objects on the table require visual markers on their upper sides to distinguish them. Also, mounting a camera over the table requires a supporting post. To provide for mobility of the table, and operability with conventional card handling, we set the following design criteria for AwareTable:

1. No additional visual markers or tags are attached to the upper side.
2. During operation, no devices are attached to the paper cards.
3. The table should contain all required devices.
4. Multiple paper cards and their IDs should be detected.
5. The table should display additional data on the top surface.

To meet these criteria, we used a glass with controllable transparency as the tabletop screen material[1]. The glass looks similar to frosted glass in its normal state, but the transparency can be instantly changed by applying an electric potential, since the glass incorporates a liquid crystal layer. Figure 1 shows the configuration of our tabletop system. To utilize the characteristics of the glass material, the system can detect both the locations and IDs of multiple paper cards by capturing their images with a camera mounted under the tabletop. In the frosted glass state, the top acts as a screen that can display additional data by projection. The projector can also be located within the table. The system can recognize paper cards or objects using visual markers printed on their backs.

[1] UMU SmartScreen developed by NSG UMU Products Co. Ltd. http://www.umupro.com/

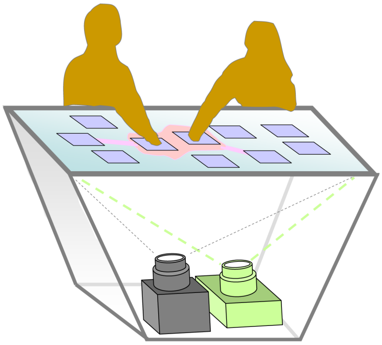


**Fig. 1.** Configuration of AwareTable. Both camera and data projector are installed in the table frame.

Therefore, the system can improve both the table's mobility and operability. By controlling the transparency of the glass, accurate scanning of visual markers and improvements in the visibility of additional projected data can be achieved.

### 2.2 Configuration

AwareTable includes (1) a glass tabletop with controlled transparency, (2) a camera for capturing visual markers on the bottom of the paper cards, and (3) a projector for displaying additional data. Digital pens store handwritten text or drawings written on the front side of the cards. The visual markers on the backs of the paper cards are printed using a standard laser printer.

### 2.3 Scenario in KJ Method Task

We describe a scenario for using the AwareTable with the KJ method[2], with multiple users. In preparation, the users print visual markers on the lower side of the paper cards. The users write their thoughts or ideas on the paper cards with a digital pen. The relationship between the paper card and the text can be either established using Anoto pens or some other technique. One solution is having the user write a number on the paper card, and letting the system recognize it. After writing, the users place the cards on the table and move them with their hands. In the KJ method, deep understanding of the text written on the card is important. However, the size of the card is insufficient to indicate its context or any background information. In such cases, AwareTable displays the context on the tabletop screen near the card. While the screen is off, the users can record the transactional process of card organization in the scanning mode. This record can be used to recall the process, or for further organizational tasks in both the physical and virtual worlds. In addition, the scanned data can be

[2] The KJ method is a registered trademark of the Kawakita research institute.

distributed to a distant place. Remote collaborative work is possible if we use a pair of tables.

A grouping operation in the KJ method can be realized by introducing special folding cards. The folding card, which represents a more advanced concept than an original card, is twice the size of an original card, and a visual marker is printed on its front. The user can fold the card in half and place the original cards between the folds of the card. Then, the user can continue the organizational tasks in the normal way as per the KJ method.

Drawing line operations, at the visualization stage of the KJ method, is not supported. However, once the card structure is obtained, the system may project the surrounding lines automatically.

### 2.4 Advantages

Controlling the transparency of the tabletop screen material is helpful in capturing small visual markers. Since the configuration of AwareTable is relatively simple, we can easily extend the size of the tabletop screen. Also, the resolutions of both the scanning image and projection data can be increased by adding cameras and data projectors. The resolution of the scanned image determines the size of the visual marker. Using a large tabletop screen and high-resolution images, the system can handle more paper cards. Thus, our framework is suitable for handling many cards simultaneously.

As mentioned in section 2.1, the frosted mode of the screen surely improves the visibility of the projected data. By increasing screen resolution, the user can obtain precise and detailed images of background knowledge.

## 3 Related Work

The concepts of extending paperwork on the table using computation emerged in the 1990s. DigitalDesk [1,2] by Wellner et al. and EnhancedDesk [3] by Koike et al. are representative examples. DigitalDesk captures finger operations and paper documents, and integrates them. Wellner et. al. explained that, using the DigitalDesk Calculator, the user can enter numbers and operators by pointing at items printed on a paper sheet. EnhancedDesk realized the linkage of real-world objects to virtual ones by recognizing fingers and a two-dimensional matrix code printed on books.

The designers' outpost [4] is a wall-sized tangible display that recognizes the locations of physical Post-it notes through computer vision. The outpost requires an environmental camera to capture the foreground image. Although the type of collaborative task is similar, we focus on the simplicity and mobility of the table configuration by encapsulating all the required devices in it. Interactive Station, developed by Ricoh, is a tabletop system that stores hand-written text or drawings composed using conventional white board markers on a screen. The stored hand-written text or drawings can be overlaid on electronic documents displayed on the surface. The concept and configuration of the Interactive Station

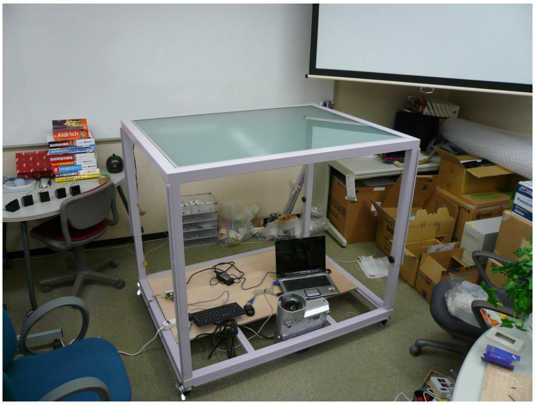


**Fig. 2.** AwareTable appearance

are similar to those of AwareTable, because both a camera and a data projector are installed in the table. However, we focus on the card-handling activity. Döring and Beckhaus proposed to apply the cards in a study on art history [5].

Some systems employ special devices to interact with digital objects in the virtual world. metaDESK [6] and Sensetable [7] are significant systems in this regard. These technologies have been applied to network management [8] and disaster simulation [9]. SnapTable [10] and PaperButtons [11] augment paper with electronic devices. SnapTable introduces electronic paper and a collaborative workspace to provide intuitive browsing of electronic documents.

Research has also been performed on screen devices using a liquid crystal shutter. Shiwa developed a large-screen telecommunication device that allows eye contact in remote conferences [12]. Kakehi et al. proposed Lumisight Table [13,14], which provides personalized projection images for four users. Lumisight Table employs Lumisty films to control the visibility of different users, and detects objects placed on the table using its inner camera. TouchLight [15] also employs a similar material for a screen. The Lumisight table uses four Lumisty films and a Fresnel lens to improve image quality. Even though the Lumisight Table provides significant functionality, it is not suitable for a large tabletop. The simplicity of the AwareTable configuration suits a large tabletop, and allows the handling of many paper cards.

## 4 Implementation

In this section, we describe implementation of the AwareTable prototype. Figure 2 shows the table. We employed UMU Smart Screen for the tabletop material. The size of the tabletop is 60 inches diagonally (1219×914mm). The table size is determined by considering the simultaneous placement of 100 paper cards, with up to six users collaborating. To control tabletop transparency, we used a solid-state relay (Phototriac BTA24-600CWRG) for switching 100 volts power

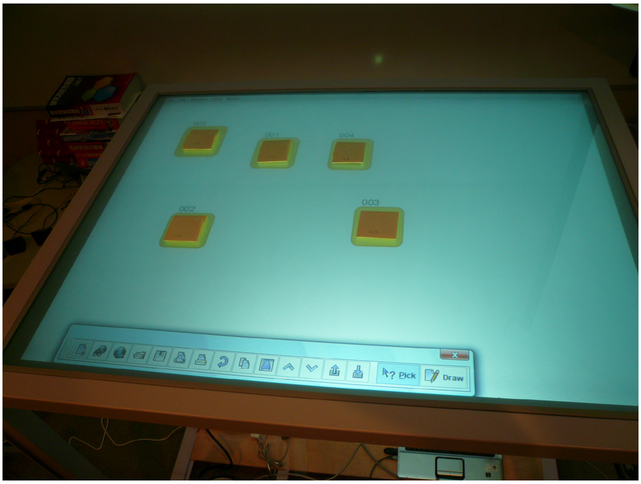


**Fig. 3.** Projection in frosted screen mode

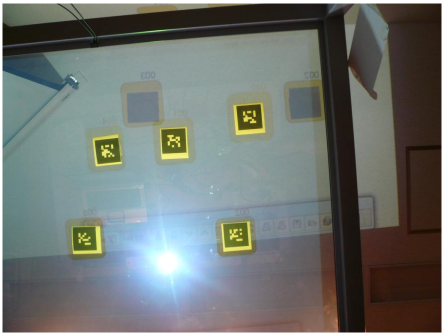


**Fig. 4.** Sample Captured Image (transparent/capture mode)

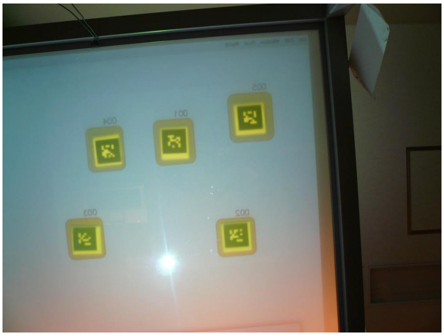


**Fig. 5.** Sample Captured Image (frosted/screen mode)

supply. The solid-state relay is controlled by a Phidgets Interface Kit[3]. For image capturing, we chose an IEEE1394 high-resolution camera (PGR Scorpion SCOR-20SOC-KT, 1600×1200 pixels). For data projection, we selected a short-throw projector (SANYO LP-XL40(S)). Figure 3 shows the image of a projection in the frosted mode.

To identify and obtain the location of each paper card, we utilized ARToolkitPlus [16] and its visual markers. ARToolkitPlus has the advantage of recognizing many visual markers with little computation. The recognition module that processes captured images from the IEEE1394 camera is constructed in Visual C++. The recognition module uses a graphic interface module written in Java, connected through Java Native Interface (JNI). A transparency control module

[3] http://www.phidgets.com/

was also developed in Visual C++, and is connected via JNI. When the graphical interface module requests scanning, the recognition module calls back with coordinates (or a translation matrix for 3D application). Then, the graphical interface module can overlay rectangles where the visual markers are placed, or show additional data on the tabletop screen.

Figure 4 and Figure 5 show sample captured images of the transparent and frosted states, respectively. Here, we used relatively large visual markers (66mm × 66mm), and the distance from the camera to the tabletop screen was 90 cm. The former image is clearer than the latter. Thus, the system has a potential of recognizing small visual tags. We have already succeeded in recognizing smaller markers (33mm × 33mm) in the transparent mode, thus proving that our tabletop framework can handle many paper cards effectively.

## 5 Conclusion and Future Work

In this paper, we proposed AwareTable, which uses transparency controllable glass as a tabletop surface. We discussed the configuration and its merits for collaborative card-handling activities, since its scalability allows application to card-handling tasks with as much as hundred paper cards. The concept of gradually extending refers to moving physical tasks into the virtual world, with full interoperability or interactivity between these two worlds. However, even the limited time-sharing approach can extend conventional paper-based tasks without any migration to the virtual world. We consider this approach promising because some people are not willing to abandon the conventional methods that are still natural and intuitive for almost all users.

In future, we will attempt to recognize finger taps on the screen in AwareTable, to provide better interaction with additional data or other virtual objects.

## Acknowledgment

Our research is partly supported by a grant-in-aid for Scientific Research (18700117) and a fund from the Ministry of Education, Culture, Sports, Science, and Technology of Japan, under the name of "Cluster for Promotion of Science and Technology in Regional Areas."

## References

1. Wellner, P.: The DigitalDesk Calculator: Tangible Manipulation on a Desk Top Display. In: Proceedings of UIST 1991, pp. 27–33 (1991)
2. Newman, W., Wellner, P.: A Desk Supporting Computer-based Interaction with Paper Documents. In: Proceedings of CHI 1992, pp. 587–592 (1992)
3. Koike, H., Sato, Y., Kobayashi, Y.: Integrating Paper and Digital Information on EnhancedDesk: A Method for Realtime Finger Tracking on an Augmented Desk System. ACM Transactions on Computer-Human Interaction 8(4), 307–322 (2001)

4. Klemmer, S.R., Newman, M.W., Farrell, R., Bilezikjian, M., Landay, J.A.: The Designers' Outpost: A Tangible Interface for Collaborative Web Site Design. In: Proceedings of UIST 2001, pp. 1–10 (2001)
5. Döring, T., Beckhaus, S.: The Card Box at Hand: Exploring the Potentials of a Paper-Based Tangible Interface for Education and Research in Art History. In: Proceedings of the 1st international conference on Tangible and embedded interaction, pp. 87–90 (2007)
6. Ullmer, B., Ishii, H.: The metaDESK: Models and Prototype for Tangible User Interfaces. In: Proceedings of UIST 1997, pp. 223–232 (1997)
7. Patten, J., Ishii, H., Hines, J., Pangaro, G.: Sensetable: A Wireless Object Tracking Platform for Tangible User Interfaces. In: Proceedings of CHI 2001, pp. 253–260 (2001)
8. Kobayashi, K., Hirano, M., Narita, A., Ishii, H.: A Tangible Interface for IP Network Simulation. In: CHI 2003 extended abstracts, pp. 800–801 (2003)
9. Kobayashi, K., Narita, A., Hirano, M., Kase, I., Tsuchida, S., Omi, T., Kakizaki, T., Hosokawa, T.: Collaborative Simulation Interface for Planning Disaster Measures. In: CHI 2006 extended abstracts (Work-in-progress), pp. 977–982 (2006)
10. Koshimizu, M., Hayashi, N., Hirose, Y.: SnapTable: Physical Handling for Digital Documents with Electronic Paper. In: Proceedings of NordiCHI 2004, pp. 401–404 (2004)
11. Pedersen, E.R., Sokoler, T., Nelson, L.: Paperbuttons: Expanding a tangible user interface. In: Proceedings of the conference on Designing interactive systems (DIS 2000), pp. 216–223 (2000)
12. Shiwa, S.: A large-screen visual telecommunications device using a liquid-crystal screen to provide eye contact. Journal of the SID, 37–41 (1993)
13. Kakehi, Y., Iida, M., Naemura, T., Shirai, Y., Matsushita, M., Ohguro, T.: Lumisight Table: An Interactive View-Dependent Tabletop Display. IEEE Computer Graphics and Applications 25(1), 48–53 (2005)
14. Kakehi, Y., Hosomi, T., Iida, M., Naemura, T., Matsushita, M.: Transparent Tabletop Interface for Multiple Users on Lumisight Table. In: Proceedings of the First IEEE International Workshop on Horizontal Interactive Human-Computer Systems (TABLETOP 2006) (2006)
15. Wilson, A.D.: TouchLight: An Imaging Touch Screen and Display for Gesture-Based Interaction. In: Proceedings of ICMI 2004, pp. 69–76 (2004)
16. Wagner, D., Schmalstieg, D.: ARToolKitPlus for Pose Tracking on Mobile Devices. In: Proceedings of 12th Computer Vision Winter Workshop (CVWW 2007) (2007)

# Random Display of Hints and Its Effect on Generating Ideas in Brain-Writing Groupware

Ujjwal Neupane[1], Kazushi Nishimoto[2], Motoki Miura[1], and Susumu Kunifuji[1]

[1] School of Knowledge Science
[2] Center of Knowledge Science
Japan Advance Institute of Science and Technology
1-1 Asahidai, Nomi-City, Ishikawa 923-1292, Japan
{neupane,knishi,miuramo,kuni}@jaist.ac.jp
http://css.jaist.ac.jp/

**Abstract.** Reflecting on the results of previous experiments, we proposed methodology for effective visualization in Brain-writing groupware. This present research mainly explains the approaches that highlight the effects of visualization in Brain-writing, and the results of combining and associating idea-labels. We have proposed methodology to combat the problem of overload viewing by analyzing the individual's ability to combine, and associate idea-labels.We conduct a test with five different modes, which are all simple in structure and are very similar in theory. Test results shows that gradual display of hints, rather than package display generates better quality ideas by maintaining the cognitive stimulation of participants, as well as keep alive the demands of individuals to view and share all ideas within groups.

**Keywords:** Divergent thinking, Brain-Writing, Visualization of ideas, Idea combination and association.

## 1 Introduction

There are varieties of approaches and techniques to conduct idea-generation; a few are: Brainstorming, Brain-writing, and Mind mapping. But we focused on the concept of Brain-writing, as several research studies show that the process of Brain-writing is effective to generate ideas in collaboration [1]. The main aim of developing a Brain-writing support system is to enhance divergent production both in quality and quantity, by dramatically increasing the creativity of the group. And, this dramatic increase can be achieved by balancing the quantity of inputs and outputs. Previous results shows that people were happy using the system and we got some good data in Brain-writing groupware, but lack of observation regarding the numbers of visible ideas actually degenerated the overall outcome [2].

Our previous results show that synergy is not always an advantage to generate ideas in groups; at times it has negative effects, as participants will be absorbed only in reading other people's ideas, and not input new ideas of their own [2]. On the other hand, results also show that participants want to view all idea sheets

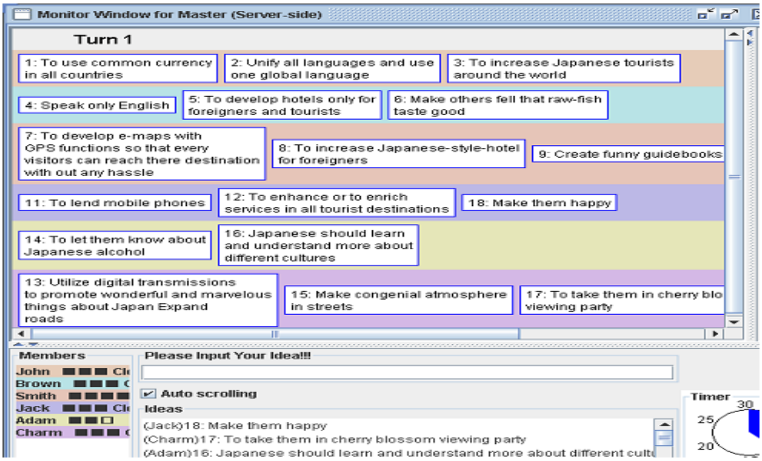
I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 863–870, 2008.

over the course of the meeting, and they acknowledged the system that supports this desire as one of the best and appropriate systems to generate ideas [2]. The comparison study of Pool Writing and Gallery Writing (these approaches are comparatively close to Brain-writing ) in which the author (s) summarized the studies conducted on 'Group Support Systems,' shows that, "meeting participants want to be able to view all comments written by group members at any given time or over the course of the meetings [3]." However, no single studies were conducted to verify the flexible numbers of idea-labels that should be displayed on the individual monitor, in accordance with appropriate time, without hampering the input process of individuals. Therefore, to clarify the numbers of ideas that should be made available for viewing; so that participants will still have time to generate newer ideas, it is necessary to study the minimum capacity of the human brain to make associations by combining idea-labels. This paper generally reflects the approach that we are going to undertake to overcome the drawbacks of information overload in idea-generation in Brain-writing groupware. Concatenate two aspects: input and sharing of ideas thus helps to design a fascinating Brain-writing support system.

## 2 Brain-Writing Groupware and Previous Insights

Brain-writing, a creative technique aimed to address the potential deficiencies of "Brainstorming," was a term coined in Germany, and "Holiger" invented the procedure of Brain-writing in 1968 [4]. The Brain-writing process can be split into two categories: traditional and automated. Compared to the traditional process of Brain-writing, the automated process is usually considered to be more versatile, as it is capable of accommodating many users and serving many functions. In face-to-face meetings some participants might be reluctant to express his/her ideas within a group, but the online distributed environment has the potential to defuse such tension, and allows groups to speak about sensitive issues in an open and candid way without the fear of judgment or shyness that characterize face-to-face groups [5] Another, potential advantage of automated meetings is that the participants gain benefits from simultaneous input/output. Moreover, the fundamental procedures of Brain-writing, in which participants input their ideas on pieces of paper and exchange ideas with others, do not need oral communication, and hence it is easy to implement in a distributed environment.

In Brain-writing, ideas generated by individuals are written down on paper, and then exchanged and combined with those of other individuals in the group. Written ideas are circulated and read by every other participant in the group, who in turn add newer ideas. In general six participants in a group generate and write three ideas in five minutes. After five minutes, in the second round, each participant passes the paper to the person on their right, who adds three more ideas [4]. This process continues until a fixed time has passed, or until each participant gets their original paper back. The paper-based traditional Brain-writing process has restrictions on input of ideas, and participants were also restricted from viewing other ideas. However, the restrictions on input of ideas and the restrictions on viewing other ideas can be suppressed technically by designing the system more specifically. Moreover, the characteristics of traditional Brain-writing make it easier to tame it in a distributed environment. Therefore, in an effort to overcome the drawbacks of traditional Brain-writing, and to improve the efficiency and effectiveness of the whole Brain-writing process, we have proposed and developed a prototype system with subsumed architecture of three different modes: [1Sheet-3+ideas] (1s-3+i), [6sheets-3+ideas] (6s-3+i), and [1Sheet-3ideas] (1s-3i); each with different functions and characteristics. The main window of the prototype is presented in Figure 1.



1: To use common currency in all countries
2: Unify all languages and use one global language
3: To increase Japanese tourists around the world
4: Speak only English
5: To develop hotels only for foreigners and tourists
6: Make others fell that raw-fish taste good
7: To develop e-maps with GPS functions so that every visitors can reach there destination with out any hassle
8: To increase Japanese-style-hotel for foreigners
9: Create funny guidebooks
11: To lend mobile phones
12: To enhance or to enrich services in all tourist destinations
18: Make them happy
14: To let them know about Japanese alcohol
16: Japanese should learn and understand more about different cultures
13: Utilize digital transmissions to promote wonderful and marvelous things about Japan Expand roads
15: Make congenial atmosphere in streets
17: To take them in cherry blo viewing party

(Jack)18: Make them happy
(Charm)17: To take them in cherry blossom viewing party
(Adam)16: Japanese should learn and understand more about different cult

**Fig. 1.** Ongoing process of Brain-writing in Prototype system

In (1s-3+i) mode, in their "first turn" participants were restricted from viewing other ideas. After all the participants input at list three ideas (minimum), they can then proceed to the "second turn", where they can view one sheet of ideas. Hereafter, in every new "turn", according to login order, the system automatically passes one new idea sheet to individuals in a round robin process. In the case of inputting ideas, in every "turn" each participant is able to input as many ideas as they can.

In (6s-3+i) mode, from the very beginning participants can view all input ideas without any restrictions. The idea which one participant inputs at his/her keyboard is immediately transmitted to the group as a whole. Participants can also input ideas without any restrictions.

(1s-3i) mode is completely based on the traditional concept of Brain-writing, where participants were restricted from viewing others' ideas at the beginning. Moreover, each participant is restricted from inputting more than three ideas in each "turn." If any participant finished inputting three ideas in a "turn", system automatically sent a message to his/her monitor to wait until all participants had input the minimum number of ideas. They have to wait for other participants to input their three ideas, and when all participants have input three ideas, they can proceed to the next "turn".

Adhering, the three modes of the prototype system we conducted test. Our past results showed that participants wanted to view all idea sheets over the course of the meeting, and they acknowledged that the system that supports this desire is one of the best and appropriate systems to generate ideas. However, the mode (6s-3+i) which permits full visibility of ideas shows negative impact as well, because participants were absorbed only to read other ideas, and not input newer ideas of their own. Without diminution or exception we received positive outcomes. However, lack of observation in regard to the components of visibility of ideas in (6s-3+i) actually degenerated the overall outcome.

## 3 Proposed Methodology and User Study

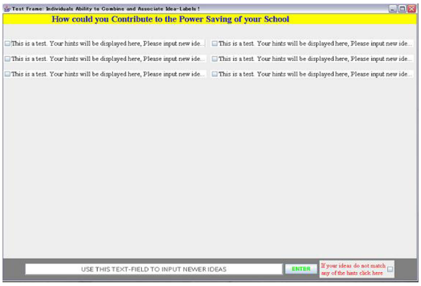
The prime objective in sharing others' ideas is that when they themselves are deadlocked in progress, they can find other ideas from which they can refine, harvest and produce new grains of ideas; literally newer ideas are the consequence of sharing others' ideas. Research shows that effective decisions or creative solution meetings, whether face-to-face or computer networked, require a full exchange of ideas by all group members [6]. But, when people have lots of ideas on the display in front of them, as they have while using (6s-3+i), they may take time to sort out ideas which they should select and associate, or which they should combine. And in this process they lose their creative momentum and therefore, could not generate large number of ideas. The ability of the human mind to combine and associate idea-labels and variations on displaying idea-labels in Brain-writing groupware will overcome the barriers, as well as production blockage. Thus, the whole process of Brain-writing can be brought to a successful end.

No work has been done regarding idea-sharing in groups, and no researchers have proposed the number of flexible ideas that should be shared in groups in accordance with appropriate timing. However, two conditions, attention (group members may not be very attentive to the ideas expressed in the group) and incubation (members may take some time to reflect on the shared information and to integrate this with their own ideas) had been proposed in which idea-sharing in groups can be productive [7]. This kind of hypothesis generally suits the traditional brainstorming paradigm, but does not match the real-time Brain-writing paradigm. According to the "cognitive stimulation theory," as long as group members pay careful attention to the shared ideas there is much potential for cognitive stimulation [8]. Although there is much potential in stimulation, participants in real time meetings were required to generate a large number of ideas within an allocated time. Therefore, they were unable to get much creative stimulation, as large number of ideas appeared instantly on their monitor, and participants finished up by only reading them. To preserve the cognitive stimulation impact, as well as the demand of individuals to view and share all ideas, we had decided to examine the minimum ability and minimum time taken by individuals to combine and associate idea-labels.

In order to clarify the desired numbers of ideas that should be made available for viewing, so that participants will still have time to generate newer ideas, we designed an individual-based idea generating tool. The necessity to verify how an individual reacts when the quantity of information on their display changes, compelled us to design a tool with five different modes: *Mode 1, Mode 2, Mode 3, Mode 4, and Mode 5* respectively. Although, the tool was composed of five different modes, the modes are simple in structure and are very similar in theory. The first mode, *Mode 1*, has a text field where ideas related to the particular theme could be input. *Mode 1* does not provide any other information, such as hints or suggestions, rather it only displays the 'theme' at the top of the display. *Mode 2*, on the other hand, provides three ideas as a hint by displaying them on the monitor. At the interval of every three minutes, the tool uploads three newer ideas by randomly selecting them from the pool of 30 ideas that were collected beforehand, and displays them on the monitor. In *Mode 3*, the initial number of hints that are displayed on the monitor is six. At the interval of every six minutes, the tool uploads six newer ideas, by randomly selecting them from the pool of 30 ideas that were collected beforehand, and displays them on the monitor. In *Mode 4*, the initial number of hints that are displayed on the monitor is 15, and 15 hints will be added in the interval of 15 minutes. On the other hand, 30 hints were displayed instantly in *Mode 5*. There are no updates of hints in *Mode 5*. Participants can view all 30 hints from the initial phase of the idea generating process to the end of the process.

Each hint displayed on the monitor has a checkbox. If subjects combined any number of hints to generate newer ideas, or if subjects associated the hint and generated newer ideas, subjects were told to click that particular hint so that their log of combination and association could be verified at the analysis stage. Subjects were free to use any number of hints at a time, and subjects were allowed to use the same hints multiple times. The system automatically stores the log of the hints clicked, and the new idea that reflects the notion of that particular hint, along with the idea input time. The image of Mode 3, and Mode 5 are presented in Figure 2 and Figure 3 respectively.

Implementing the above described modes, we conduct a comparison experiment to study the individuals' ability to combine and associate ideas, using 25


Test Frame: Individuals Ability to Combine and Associate Idea-Labels !

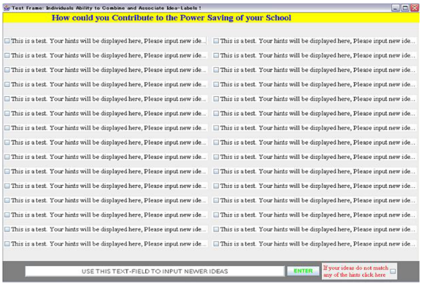
How could you Contribute to the Power Saving of your School

□ This is a test. Your hints will be displayed here, Please input new ide...
□ This is a test. Your hints will be displayed here, Please input new ide...
□ This is a test. Your hints will be displayed here, Please input new ide...
□ This is a test. Your hints will be displayed here, Please input new ide...
□ This is a test. Your hints will be displayed here, Please input new ide...
□ This is a test. Your hints will be displayed here, Please input new ide...

USE THIS TEXT-FIELD TO INPUT NEWER IDEAS

If your ideas do not match any of the hints click here □

**Fig. 2.** Display of hints in Mode 3


Test Frame: Individuals Ability to Combine and Associate Idea-Labels !

How could you Contribute to the Power Saving of your School

□ This is a test. Your hints will be displayed here, Please input new ide...
□ This is a test. Your hints will be displayed here, Please input new ide...
□ This is a test. Your hints will be displayed here, Please input new ide...
□ This is a test. Your hints will be displayed here, Please input new ide...
□ This is a test. Your hints will be displayed here, Please input new ide...
□ This is a test. Your hints will be displayed here, Please input new ide...
□ This is a test. Your hints will be displayed here, Please input new ide...
□ This is a test. Your hints will be displayed here, Please input new ide...
□ This is a test. Your hints will be displayed here, Please input new ide...
□ This is a test. Your hints will be displayed here, Please input new ide...
□ This is a test. Your hints will be displayed here, Please input new ide...
□ This is a test. Your hints will be displayed here, Please input new ide...
□ This is a test. Your hints will be displayed here, Please input new ide...
□ This is a test. Your hints will be displayed here, Please input new ide...
□ This is a test. Your hints will be displayed here, Please input new ide...
□ This is a test. Your hints will be displayed here, Please input new ide...
□ This is a test. Your hints will be displayed here, Please input new ide...
□ This is a test. Your hints will be displayed here, Please input new ide...
□ This is a test. Your hints will be displayed here, Please input new ide...
□ This is a test. Your hints will be displayed here, Please input new ide...
□ This is a test. Your hints will be displayed here, Please input new ide...
□ This is a test. Your hints will be displayed here, Please input new ide...
□ This is a test. Your hints will be displayed here, Please input new ide...
□ This is a test. Your hints will be displayed here, Please input new ide...
□ This is a test. Your hints will be displayed here, Please input new ide...
□ This is a test. Your hints will be displayed here, Please input new ide...
□ This is a test. Your hints will be displayed here, Please input new ide...
□ This is a test. Your hints will be displayed here, Please input new ide...
□ This is a test. Your hints will be displayed here, Please input new ide...
□ This is a test. Your hints will be displayed here, Please input new ide...

USE THIS TEXT-FIELD TO INPUT NEWER IDEAS

If your ideas do not match any of the hints click here □

**Fig. 3.** Display of hints in Mode 5

subjects. The subjects were randomly divided into five different groups from group one to group five and each subject in the five groups experienced the test individually. The *first group* undergoes the test using *Mode 1*. The *second group* took part in the test using *Mode 2*. The *third, fourth, and fifth group*, took part in the test by using *Mode 3*, *Mode 4*, and *Mode 5* respectively. Due to the differences in the number of visible hints in each mode, the exposure time per hint in each mode was also different. For instance, in *Mode 5*, 30 hints were displayed instantly. Therefore, the volume of exposure is (30 hints × 30 minutes) 900 hints per subject. On the other hand subjects experiencing *Mode 2* had the exposure rate of (3 hints × 30 minutes + 3 hints × 27 minutes + $\cdots$ + 3 hints × 3 minutes) 495 hints. Ultimately, this makes a big difference in the volume of exposure. To avoid the differences, we decided to standardize the volume of exposure based on the exposure rate of 495 hints while using *Mode 2*. Therefore, for *Mode 1* and *Mode 2*, we allocated 30 minutes. We allocated 27.5 minutes for *Mode 3*, 22 minutes for *Mode 4* and 16.5 minutes for *Mode 5* respectively.

Regardless of mode, to provide similar opportunity for all subjects at the beginning of the test, we allocated five minutes. In the first allocated five minutes all subjects used the same function where they were not provided with any hints. On all of the tests, subjects were told to think about "how they could contribute to power saving at their school."

## 4 Results

The total number of qualitative ideas and quantitative ideas generated by each subjects is presented in Table 1.

For the qualitative evaluation of ideas, we asked three individuals to evaluate the ideas based on a five stage rating system. After the evaluation, the average of each idea was calculated and the ideas with more than three points on average were considered to be good ideas. Based on the log file, we verified the patterns of combined ideas, associated ideas and directly-input ideas. The numbers of combined, associated, and directly input ideas are presented in Table 2.

**Table 1.** Total number of ideas (Left),Total number of qualitative ideas (Right)

| Subject | S:1 | S:2 | S:3 | S:4 | S:5 | Total | Average |
|---|---|---|---|---|---|---|---|
| Mode 1 | 34 | 25 | 120 | 21 | 37 | 237 | 47 |
| Mode 2 | 17 | 24 | 49 | 60 | 19 | 169 | 34 |
| Mode 3 | 19 | 26 | 21 | 26 | 41 | 133 | 27 |
| Mode 4 | 27 | 23 | 14 | 24 | 29 | 117 | 23 |
| Mode 5 | 11 | 35 | 9 | 7 | 17 | 79 | 16 |

| Subjects | | S:1 | S:2 | S:3 | S:4 | S:5 |
|---|---|---|---|---|---|---|
| Mode 1 | First 5 minutes | 3 | 1 | 2 | 3 | 2 |
| | After 5 minutes | 15 | 13 | 37 | 3 | 5 |
| Mode 2 | First 5 minutes | 5 | 4 | 6 | 5 | 10 |
| | After 5 minutes | 6 | 11 | 20 | 6 | 0 |
| Mode 3 | First 5 minutes | 4 | 2 | 5 | 5 | 8 |
| | After 5 minutes | 9 | 11 | 9 | 10 | 3 |
| Mode 4 | First 5 minutes | 7 | 4 | 4 | 4 | 7 |
| | After 5 minutes | 2 | 5 | 4 | 10 | 12 |
| Mode 5 | First 5 minutes | 2 | 5 | 2 | 3 | 4 |
| | After 5 minutes | 2 | 10 | 0 | 1 | 5 |

**Table 2.** Numbers of Combined, Associated and Direct Ideas

| Subjects | | S:1 | S:2 | S:3 | S:4 | S:5 | Total |
|---|---|---|---|---|---|---|---|
| | Combination | 1 | 12 | 11 | 0 | 2 | 26 |
| Mode 2 | Association | 8 | 7 | 19 | 14 | 5 | 53 |
| | Direct | 3 | 0 | 6 | 32 | 0 | 41 |
| Mode 3 | Combination | 2 | 3 | 9 | 8 | 2 | 24 |
| | Association | 49 | 12 | 5 | 10 | 10 | 41 |
| | Direct | 9 | 5 | 1 | 1 | 19 | 35 |
| Mode 4 | Combination | 4 | 5 | 1 | 0 | 13 | 23 |
| | Association | 14 | 9 | 6 | 13 | 6 | 48 |
| | Direct | 1 | 0 | 2 | 7 | 1 | 11 |
| Mode 5 | Combination | 2 | 3 | 0 | 1 | 1 | 7 |
| | Association | 2 | 15 | 0 | 2 | 13 | 22 |
| | Direct | 2 | 3 | 7 | 0 | 6 | 18 |

We asked subjects their impression of the Modes by conducting a questionnaire session subsequently. The answers of the questionnaire were evaluated by adopting a seven stage rating system. The average values of the answer to the question “Did you read all displayed hints throughout the session?” were different for each mode: six for Mode 2, six for Mode 3, seven for Mode 4 and five for Mode 5 respectively. For the question, “Was the time sufficient to read all hints and generate newer ideas?”, the average values of the answer were five for Mode 2, five for Mode 3, four for Mode 4 and three for Mode 5 respectively. We asked them “Could you have generated a large number of ideas if you were provided more hints?”. The average values of the answer to the question were, four for Mode 2, three for Mode 3, four for Mode 4, and three for Mode 5 respectively. Furthermore, all participants experiencing Mode 1, where they do not receive any hints replied that they felt it was difficult to generate ideas, and that they would have generated more ideas if hints were provided to them.

## 5 Conclusions

In this paper we proposed methodology to combat the problem of overload viewing by analyzing the minimum time and minimum number of idea that should be made visible and sharable with participants in a way that still maintains the cognitive stimulation of participants, as well as meet the demands of individuals to view and share all ideas.

In each mode subjects who generated more ideas in the first five minutes also generated more ideas at the later stage. On the other hand, it was observed that subjects who had fewer ideas in the first five minutes generated a larger number of ideas after they were stimulated by hints. The pattern of displaying hints does not matter much to subjects who are good at divergent thinking, as they demonstrate continuous increase in generating ideas. But all patterns of hints display that they support the idea generating process of subjects with fewer ideas by stimulating them to generate a large number of ideas. They tended to combine and associate hints to increase the volume of ideas.

On the qualitative evaluation of ideas, the percentage of good ideas of modes where hints are provided exceeds the results of Mode 1 where subjects do not receive any hints. However, due to fewer samples, the explicit differences among the modes where hints are provided were not observed. Nevertheless, intimate relationship between the qualitative ideas and the patterns of combination and association were observed in all modes. Almost all qualitative ideas were generated by either associating or combining hints. Moreover it was observed that, subjects combine more pairs of hints when hints are provided gradually rather than at times when large numbers of hints were displayed instantly. This may be because, when larger volumes of hints are around, subjects may get mired in confusion, and thus be unable to decide which pairs of ideas could be combined.

Our results shows that gradual pop-up display of hints, rather than package display, generates better quality ideas, as well as meet the demands of individuals to view and share all ideas.

## References

1. Davison, R.M., Briggs, R.O.: Gss for presentation support. Communications of the ACM 43(9), 91–97 (2000)
2. Neupane, U., Miura, M., Hayama, T., Kunifuji, S.: Qualitative, quantitative evaluation of ideas in brain writing groupware. The Institute of Electronics, Information and Communication Engineers E90-D(10), 1493–1499 (2007)
3. Aiken, M., Vanjani, M.B.: Comment distribution in electronic poolwriting and gallery writing meetings. Communications of the International Information Management Association 3(2)
4. Takahashi, M.: The bible of creativity, pp. 294–296. JUSE press. Ltd (2002)
5. Sweet, C.: Expanding the qualitative research arena: Online focus groups. Quesst Qualitative Research, New York (1999)
6. Janis, I.L., Mann, L.: Decision making: A psychological analysis of conflict, choice, and commitment. Free Press, New York (1997)
7. Paulus, P.B., Yang, H.C.: Dea generation in groups: A basis for creativity in organizations. Organizational behavior and human decision process 82(1), 76–87 (2000)
8. Paulus, P.B.: Groups, teams and creativity: The creative potential of idea generating groups. Applied Psychology: An International Review 49, 237–262 (2000)

# Visual Analysis Tool for Bipartite Networks

Kazuo Misue

Department of Computer Science, University of Tsukuba,
1-1-1 Tennoudai, Tsukuba, 305-8573 Japan
misue@cs.tsukuba.ac.jp

**Abstract.** To find hidden features of co-authoring relationships, we focused on drawing techniques for bipartite graphs. We previously developed a drawing method called "anchored maps", which is used to depict bipartite graphs visually. In anchored maps, some nodes are restricted to certain positions, but others are left to be arranged freely. In this study, we used our method to depict co-authoring relationships as anchored maps. The maps revealed certain co-authoring relationships. Anchored maps can help us discover important features that we cannot find when using only bipartite graphs that have a free layout.

**Keywords:** information visualization, graph drawing, bipartite graph, anchored map, co-authorship network.

## 1 Introduction

Co-authoring relationships are represented as an undirected graph, but its structure is based on a bipartite graph that consists of academic papers and their authors. There are some research results related to drawing techniques of bipartite graphs[4,13,12,2]. However, most of them put their focus on theoretical aspects, such as planar layout and edge-crossing minimization. Drawing techniques for bipartite graphs focus on applications have not been studied very much. To find hidden features of co-authoring relationships visually, drawing techniques that are more application oriented are desired.

We developed a method of drawing bipartite graphs called "anchored map[8,9]." An anchored map restricts some nodes to certain positions but leaves other nodes so they can be arranged freely. Some examples of anchored maps can be found in published works[10]. Such examples illustrate what kind of knowledge users can find from anchored maps.

This paper describes a visual analysis tool that has two main functions: constructing bipartite networks that represent relationships between academic papers and their authors, and drawing bipartite networks as anchored maps. We took relationships between academic papers and their authors as examples and investigated the insight given into the relationships by anchored maps.

## 2 Academic Papers/Authors Networks

We focused on networks that can be formalized as bipartite graphs. A bipartite graph is produced by representing papers and authors by nodes, and then

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 871–878, 2008.

connecting by edges every node representing a paper to the corresponding nodes representing the paper's author(s).

### 2.1 Formal Description of Bipartite Networks

The node set of a bipartite graph is divided into two exclusive sets. A bipartite graph is represented as $G = (A \cup B, E)$. Here, $A$ and $B$ are finite sets of nodes, and $A$ and $B$ are disjoint. $E$ is a finite set of edges, and $E$ is a subset of $A \times B$.

### 2.2 Network Source

We used DBLP[1] to get real data. The DBLP, which is maintained by Michael Ley, is a database server providing bibliographic information on major computer science journals and proceedings. The server indexed about 950,000 articles on October 29, 2007. Each record at DBLP corresponds to a paper and contains its title, author names, conference name, journal name, page numbers, published year, etc. We extracted title, author names and published year of every paper and then stored them in an RDB for ease of construction of local structures.

### 2.3 How to Extract Local Structures

Visualization of the whole of the DBLP is a challenging task, but we focused on the more detailed structures of the DBLP. We decided to perform a breadth-first search beginning with an author (or a paper) in order to extract a local structure of the database.

The notations used in our searches are as follows. A set of all the papers written by author $a$ is denoted by $P(a)$. A set of all the co-authors who wrote paper $p$ is denoted by $A(p)$. Furthermore, a set of all the papers written by authors in a set $A$ is expressed with $P(A)$, that is $P(A) = \{p \in P(a) | a \in A\}$. A set of all the co-authors who wrote papers in a set $P$ is expressed with $A(P)$, that is $A(P) = \{a \in A(p) | p \in P\}$.

A search that begins with author $a_0$ is performed as follows. All papers written by author $a_0$ are searched for initially. We assumed that the set of all the papers written by author $a_0$ is $P_0$, that is $P_0 = P(a_0)$. Then, we search for all the co-authors of all the papers in the set $P_0$. Let $A_1$ denote the set of all the found co-authors, that is $A_1 = A(P_0)$. Next, we search for all the papers written by all the co-authors in $A_1$ . Let $P_1$ denote the set of all the found papers, that is $P_1 = P(A_1)$. We obtain a bipartite graph described as follows.

$$G_1 = (A_1 \cup P_1, E_1) \tag{1}$$

$$E_1 = \{(a, p) \in A_1 \times P_1 | a \in A(p)\} \tag{2}$$

It seems that to get $P_0$ and $A_1$ as a search depth of 1 is adequate. However, we perform the search until $P_1$ because we want to find joint publications by some of the co-authors in $A_1 \setminus \{a_0\}$.

[1] http://dblp.uni-trier.de

We obtain the following bipartite graph $G_k$ by repeating the search $k$ ($k = 1, 2, \ldots$) times (called "search of $k$ in depth"):

$$G_k = (A_k \cup P_k, E_k) \tag{3}$$

$$E_k = \{(a, p) \in A_k \times P_k | a \in A(p)\} \tag{4}$$

Where $A_0 = \{a_0\}$, $P_i = P(A_i)$, $A_{i+1} = A(P_i)$ ($i = 0, \ldots, k$).

# 3 Visualization of Bipartite Networks

## 3.1 Free Layout by Spring Embedder Model

Eades' spring embedder model[3] is a well-known method for drawing undirected graphs. We can draw bipartite graphs by using a spring embedder model if we ignore any distinctions between the two sets of nodes. One easy way is to assign different colors and/or shapes to the two sets of nodes. The problem is this produces bipartite graphs that do not have good readability.

## 3.2 Anchored Map

We developed a method of drawing bipartite graphs called "anchored map[8]." The node set of a bipartite graph is divided into two exclusive sets. Here, the nodes in one set are called "anchors," and the nodes in the other set are called "free nodes." Anchors are arranged on the circumference and free nodes are arranged at suitable positions in relation to the adjacent anchors. The anchors are fixed on the circumference as their name suggests and have an effect similar to a coordinate system in that they increase the readability of the network diagrams.

On anchored maps, the order of anchors has a significant effect on the readability of the graph structures. We examined some methods in order to decide the order of anchors[9].

# 4 Implementation of Tool

## 4.1 Service of Papers/Authors Graph

We developed a program to perform the search described in Section 2. The program accepts not only search depth but also a minimum degree of author nodes and a minimum degree of paper nodes. The number of the nodes often increases exponentially when we perform a deep search[2]. There are a lot of authors who wrote only one or two papers with their supervisor when they were college students. Depending on the purpose, we can ignore such authors with a small of number of published papers. Our tool can specify the smallest

[2] It does not always increase. When joint works are performed in a closed community, the author nodes do not increase even if the search is deepened.

degree of the author nodes, and remove the author nodes with a smaller degree than the specified degree. The smallest degree of the paper nodes can also be specified. For example, papers written by only one author do not affect the social network consisting of co-authoring relationships. When we are interested in a co-authorship network, we can remove such papers by setting the smallest degree of paper nodes to be 2.

The program has been developed as a web service and it can be accessed by the public. The service returns bipartite graphs in the graphML format[1].

### 4.2 Visualization Tool Service

We also developed a Java applet that provides facilities for drawing anchored maps. The applet also provides facilities for deciding anchor order according to several criteria. Furthermore, we provide a web page[3] with some forms for specifying parameters in order to invoke the Java applet. On the web page, enter the name of an author, search depth, minimum degree of author nodes, and minimum degree of paper nodes. Clicking a button invokes the Java applet. Furthermore, we can obtain anchored maps by selecting the function from a menu.

## 5 Examples and Discussion

When we designated an author name “Kazuo Misue” (the author of this paper) as the starting point and 2 as the depth of search on the web page and then clicked “Draw Network” button, a bipartite graph was provided. The graph had 117 author nodes, 627 paper nodes, and 1,500 edges.

### 5.1 View as a Free Layout

Figure 1 shows the network drawn using the spring embedder model. A rectangle represents an author node and a bullet point represents a paper node. The colors of the paper nodes represent the publishing year of the paper (but it might be difficult to distinguish the colors in the Proceedings). Blue means old and red means new. Author nodes that have a lot of red nodes represent currently active communities. For example, the authors in the vicinity of the central right side of the figure are more active than the authors in the lower left of the figure.

### 5.2 Authors Are Fixed as Anchors

Figure 2(a) shows the network drawn as an anchored map. Here, the author nodes are fixed as anchors. The order of author nodes (anchors) has been decided so as to make author nodes that are adjacent to common paper nodes (free nodes) close to each other. If every paper is written by authors in a local community, co-authors of a paper are placed close to each other and the paper comes close

[3] http://www.iplab.cs.tsukuba.ac.jp/˜misue/open/dblp2/

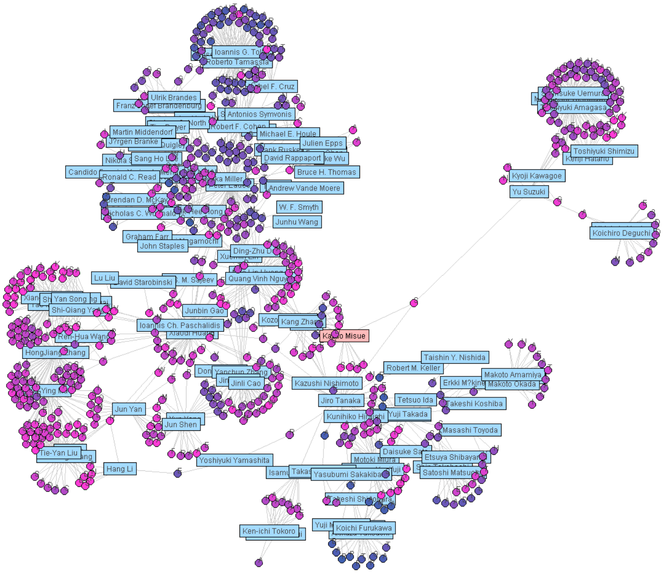


**Fig. 1.** Examples of a papers/authors network drawn using spring embedder model. Network consists of 117 author nodes (rectangles), 627 paper nodes (small circles), and 1500 edges.

to the co-authors, that is, the circumference. However, some paper nodes are placed far from the circumference. This means that such papers may have been written by authors from different communities.

For instance, the paper[4] placed near the center of Figure 2(a) was written by four co-authors. When they wrote the paper, two of them worked at a private company and other two were at a university. They moved to different universities and continued research activities afterwards. From the viewpoint of their current positions, this paper is crossing communities. Anchored map helps us discover such papers.

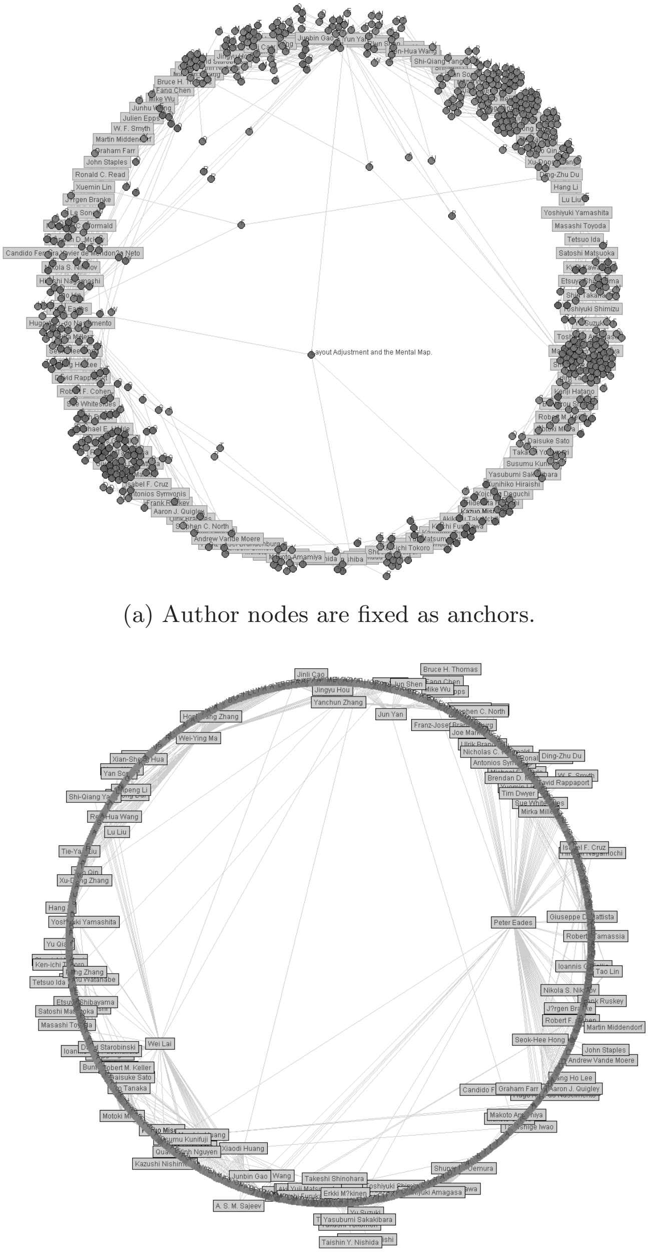
### 5.3 Papers Are Fixed as Anchors

Figure 2(b) is the same network as Figure 2(a). Here, paper nodes are fixed as anchors; in other words in Figure 2(b) anchors and free nodes are exchanged with those in Figure 2(a). The order of paper nodes (anchors) has been decided in order to make paper nodes that are adjacent to common author nodes (free nodes) close to each other. Nodes of authors who wrote a lot of papers with various co-authors are toward the center of the circle. For example, we see such a node in the central right side[5] of the figure and in the lower left[6] of the figure.

[4] K. Misue, P. Eades, W. Lai and K. Sugiyama: Layout Adjustment and the Mental Map, Journal of Visual Languages and Computing, Vol. 6, No. 2, pp. 183–210, 1995.

[5] Peter Eades, the author of [3].

[6] Wei Lai.



(a) Author nodes are fixed as anchors.

(b) Paper nodes are fixed as anchors.

**Fig. 2.** Examples of a papers/authors network drawn as anchored maps. Network is same as in Figure 1.

Furthermore, we can assume that the two authors do not have many joint papers because the two nodes are placed apart from each other.

### 5.4 Discussion

The anchored maps are an effective tool for finding distinctive nodes. They can help us discover important features that we cannot find with feature quantities such as the degree of nodes. It is possible to obtain an overview of graph structures with free layout by the spring embedder model, but it is difficult to find distinctive structures. Actually, it is difficult to find the characteristic paper nodes and author nodes mentioned above only from Figure 1.

Various studies have been conducted in order to analyze co-authorship networks (e.g., [11,7]). Some researchers have drawn co-authorship networks as diagrams, as explained in Section3.1[6,7]. In the diagrams, attributes such as the number of papers, number of citations, and number of times co-authored, are represented by giving visual attributes (e.g., color, size and width) to the nodes and edges. The local characteristics of the networks are expressed visually. However, the structural characteristics are not always expressed very well.

Huang et al. developed InterRing, which visualizes co-authorship without network diagrams[5]. In InterRing, change of co-authors' contributions are represented in a pie-chart style. It helps us see a researcher's collaboration history in a given period; however, it cannot help us discover some of the structural characteristics.

## 6 Concluding Remarks

We developed a tool to visualize papers/authors networks as anchored maps. From observing drawing examples, it can be seen that anchored maps are an effective way of finding characteristic components in networks. More work is needed to bring out advantages of anchored map; We will apply anchored maps to other kind of networks, and will extend the style of anchored maps.

## References

1. Brandes, U., Marshall, M.S., North, S.C.: Graph data format workshop report. In: Marks, J. (ed.) GD 2000. LNCS, vol. 1984, pp. 407–409. Springer, Heidelberg (2001)
2. Di Giacomo, E., Grilli, L., Liotta, G.: Drawing Bipartite Graphs on Two Curves. In: Kaufmann, M., Wagner, D. (eds.) GD 2006. LNCS, vol. 4372, pp. 380–385. Springer, Heidelberg (2007)
3. Eades, P.: A Heuristic for Graph Drawing. Congressus Numerantium 42, 149–160 (1984)
4. Fosmeier, U., Kaufmann, M.: Nice Drawings for Planar Bipartite Graphs. In: Proceedings of the 3rd Italian Conference on Algorithms and Complexity, pp. 122–134 (1997)

5. Huang, T.-H., Huang, M.L.: Analysis and Visualization of Co-authorship Networks for Understanding Academic Collaboration and Knowledge Domain of Individual Researchers. In: Proc of IEEE Computer Graphics, Imaging and Visualisation 2006 (CGIV 2006), pp. 18–23 (2006)
6. Ke, W., Börner, K., Viswanath, L.: Major Information Visualization Authors, Papers and Topics in the ACM Library. In: IEEE Symposium on Information Visualization (InfoVis 2004) (2004)
7. LaRowe, G., Ichise, R., Börner, K.: Analysis of Japanese Information Systems Coauthorship Data. In: Proceedings of 11th International Conference Information Visualization (IV 2007), pp. 459–464 (2007)
8. Misue, K.: Drawing Bipartite Graphs as Anchored Maps. In: Proceedings of Asia-Pacific Symposium on Information Visualization (APVIS 2006). CRIPT, vol. 60, pp. 169–177 (2006)
9. Misue, K.: Anchored Maps: Visualization Techniques for Drawing Bipartite Graphs. In: Jacko, J.A. (ed.) HCI 2007. LNCS, vol. 4551, pp. 106–114. Springer, Heidelberg (2007)
10. Misue, K.: Overview of Network Intormation by Using Anchored Maps. In: Apolloni, B., Howlett, R.J., Jain, L. (eds.) KES 2007, Part II. LNCS (LNAI), vol. 4693, pp. 1269–1276. Springer, Heidelberg (2007)
11. Newman, M.E.J.: Coauthorship Networks and Patterns of Scientific Collaboration. Proceedings of the National Academy of Science 101, 5200–5205 (2004)
12. Newton, M., Sykora, O., Vrto, I.: Two New Heuristics for Two-Sided Bipartite Graph Drawing. In: Goodrich, M.T., Kobourov, S.G. (eds.) GD 2002. LNCS, vol. 2528, pp. 312–319. Springer, Heidelberg (2002)
13. Zheng, L., Song, L., Eades, P.: Crossing Minimisation Problems of Drawing Bipartite Graphs in Two Clusters. In: Proceedings of Asia-Pacific Symposium on Information Visualization (APVIS 2005). CRIPT, vol. 45, pp. 33–38 (2005)

# A Model for Measuring Cognitive Complexity of Software

Sanjay Misra and Ibrahim Akman

Department of Computer Engineering, Faculty of Engineering
Atilim University, Ankara, Turkey
smisra@atilim.edu.tr, akman@atilim.edu.tr

**Abstract.** This paper proposes a model for calculating cognitive complexity of a code. This model considers all major factors responsible for (cognitive) complexity. The practical applicability of the measure is evaluated through experimentation, test cases and comparative study.

**Keywords:** Software complexity, metric, size, structure, cognitive complexity, understandability.

## 1 Introduction

Software metrics have always been important for software engineers to assure software quality because they provide approaches to the quantification of quality aspects of software. However, absolute measures are uncommon in software engineering [9]. Instead, software engineers attempt to derive a set of indirect measures that lead to metrics that provide an indication of quality of some representation of software. The quality objectives may be listed as performance, reliability, availability and maintainability [10] and are closely related to software complexity. Complexity is defined by IEEE [3] as "the degree to which a system or component has a design or implementation that is difficult to understand and verify" Over the years, research on measuring the software complexity has been carried out to understand, what makes software products difficult to develop, maintain, or use. Major complexity measures of software that refer to effort, time and memory expended have been used in the form of different software metrics. Cyclomatic number [4], Halstead programming effort [2], data flow complexity measures [8], cognitive functional size measure [11], are examples to such metrics. Number of metrics can also be found at [7]. These metrics calculate the complexity of software from the code and measures only specific internal attributes like size, algorithm complexity, control flow structures etc. In all above mentioned complexity metrics, they attempt to quantify the primitives which make software difficult to understand. For many of them, the developer's claim that their complexity metric based on an internal attribute is the most accurate predictor of software quality. However, the authors realize that a single internal attribute is not sufficient for measuring the complexity of the code. For measuring the complexity of a code, one must consider most of the internal attributes responsible for complexity. Therefore, the purpose of this paper is to propose a new complexity metric which

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 879–886, 2008.

calculates complexity of the program code by considering all factors responsible for complexity. For this, first we identified the factors which are responsible for the complexity and then established a metric to reflect a proper relationship between these factors. In our previous work, we presented a metric in ICCI, 2007, [6] which is based on input, output and basic control structures (based on cognitive informatics [12]). In the present work, we extended our previous work by including all the factors responsible for complexity of software.

In section 2, we identified the primitives responsible for the complexity and accordingly proposed a new measure. The metric is demonstrated in section 3. Experimentation and comparative study are given in section 4. The last section 5 includes the conclusions drawn.

## 2 Proposed Metric: Unified Complexity Measure (UCM)

Complexity of a code is directly dependent on the understandability of the code and relates to ease of comprehension. It is a cognitive process. All the factors that makes program difficult to understand are responsible for cognitive complexity. When we analyze a program code we find that that number of lines (size), total occurrence of operators and operands (size), numbers of control structures (control flow structuredness), function call (coupling) are the factors which directly affect the complexity. In general, these primitives are measured independently by different complexity measures and each one of these is assumed to represent overall complexity of the software. When we look at most of the known complexity measures, we can observe the close relation between number of lines, operator and operand counts, and basic control structures. Consequently, these primitives of software may constitute the components of a unified, comprehensive complexity measure.

In our opinion, the complexity of a software system depends on following factors:

1. Complexity of program depends on the size of the code. We suggest that the size of the code can be measured by total occurrence of operators and operands. Therefore, the complexity due to $i^{th}$ line of the code can be calculated as

   $$SOO_i = N_{i1} + N_{i2} \text{ Where}$$

   $N_{i1}$: The total number of occurrences of operators at line i,
   $N_{i2}$: The total number of occurrences of operands at line i,

2. Complexity of the program is directly proportional to the cognitive weights of Basic Control Structures (BSC). Cognitive weight of software [11] is the extent of difficulty or relative time and effort for comprehending given software modeled by a number of BCS's. BCS's, sequence, branch and iteration [11] are basic logic building blocks of any software and their weights are one, two and three respectively. These weights are assigned on the classification of cognitive phenomenon as discussed by Wang [11]. He proved and assigned the weights for sub conscious function, meta cognitive function and higher cognitive function as 1, 2 and 3 respectively. In fact, cognitive weights correspond to the number of executed instructions. The details of the weights for different BCS's are given in Table-1, see [11].

**Table 1.** Basic Control Structures and their weights

| Category | Basic<br>Control Structures | Cognitive<br>Weight |
|---|---|---|
| Sequence | Sequence | 1 |
| Branch | If-Then-Else | 2 |
| | Case | 3 |
| Iteration | For-do | 3 |
| | Repeat-until | 3 |
| | While-do | 3 |
| Embedded Component | Function Call | 2 |
| | Recursion | 3 |

As a result, the cognitive complexity due to i$^{th}$ line of the code, *CW$_i$,* can be weighted as in Table-1.

Using the above considerations, we propose the following model to establish a proper relationship among internal attributes of software.

$$UnifiedComplexityMeasure(UCM) = \sum_{i=1}^{n} \sum_{j=1}^{m_i} (SOO_{ij} * CW_{ij}) \qquad (1)$$

where complexity measure of the software code UCM is defined as the sum of complexity of its n modules and module i consists of $m_i$ line of code.
It is important to note here that in this formula:

- number of lines ($m_i$), number of operators and operands correspond to size of software,
- total occurrence of basic control structures, operators and operands ($SOO_{ij}$) is related to algorithm complexity,
- basic control structures ($CW_{ij}$) are related to control flow structuredness, therefore corresponds to structural complexity,
- $CW_{ij}$ also corresponds to cognitive complexity.
- number of modules (n) is related to modularity,
- function calls in terms of basic control structures are related to coupling between modules( in terms of $CW_{ij}$'s).

We believe that these are the major factors which are responsible for the program comprehension, therefore complexity of the software system.

In our context, the concept of cognitive weights is used as an integer multiplier. Therefore, the unit of the UCM (Unified Complexity Unit-UCU) is always a positive integer number. This implies achievement of scale compatibility of *SOO* and *CW.*

## 3 Demonstration of UCM

The proposed complexity metric given by equation 1 is demonstrated with the programming example given by the following Table 2.

**Table 2.** Calculated complexity values for the example program

| Line No. | Sample Algorithm | Components | | UCM$_i$ |
|---|---|---|---|---|
| | | SOO$_i$ | CW$_i$ | |
| Line 1 | #include<stdio.h> | 0 | 1 | 0 |
| Line 2 | #include<stdlib.h> | 0 | 1 | 0 |
| Line 3 | #include<conio.h> | 0 | 1 | 0 |
| Line 4 | int main (){ | 0 | 1 | 1 |
| Line 5 | long fact (int n); | 3 | 1 | 3 |
| Line 6 | int isprime(int n); | 3 | 1 | 3 |
| Line 7 | int n; | 2 | 1 | 2 |
| Line 8 | long int temp; | 2 | 1 | 2 |
| Line 9 | clrscr(); | 1 | 1 | 1 |
| Line 10 | printf("\n input the number"); | 1 | 1 | 1 |
| Line 11 | scanf("%d",&n); | 2 | 1 | 2 |
| Line 12 | temp=fact(n); | 5 | 2 | 10 |
| Line 13 | {printf("\n is prime");} | 1 | 1 | 1 |
| Line 14 | int flag1=isprime(n); | 5 | 2 | 10 |
| Line 15 | if (flag1==1) | 3 | 2 | 6 |
| Line 16 | else | 0 | 1 | 0 |
| Line 17 | {printf("\n is not prime")}; | 1 | 1 | 1 |
| Line 18 | printf("\nfactorial(n)=%d", temp); | 2 | 1 | 2 |
| Line 19 | getch(); | 1 | 1 | 1 |
| Line 20 | long fact(int n) | 2 | 1 | 2 |
| Line 21 | {long int facto=1; | 4 | 1 | 4 |
| Line 22 | if (n==0) | 3 | 2 | 6 |
| Line 23 | facto=1;else | 4 | 1 | 4 |
| Line 24 | facto=n*fact(n-1); | 9 | 1 | 9 |
| Line 25 | return (facto); } | 2 | 1 | 1 |
| Line 26 | int isprime(int n) | 2 | 1 | 2 |
| Line 27 | { int flag; | 2 | 1 | 1 |
| Line 28 | if (n==2) | 3 | 2 | 6 |
| Line 29 | flag=1; | 4 | 1 | 4 |
| Line 30 | else | 0 | 1 | 0 |
| Line 31 | for (int i=2;i<n;i++) | 10 | 3 | 30 |
| Line 32 | { if (n%i==0) | 5 | 2 | 10 |
| Line 33 | { flag=0; | 4 | 1 | 4 |
| Line 34 | break; } | 1 | 1 | 1 |
| Line 35 | else { | 0 | 1 | 0 |
| Line 36 | flag=1 ;}} | 4 | 1 | 4 |
| Line 37 | return (flag);}} | 2 | 1 | 2 |
| | TOTAL | | | 136 |

This example consists of a simple source code, which contains a main program and two functions. The main program (lines 1-19) calls the function fact (lines 20-25) to calculate the factorial of the inputted positive integer and calls the function prime (lines 26-37) to check whether the inputted integer is a prime number or not. The last three columns of table 2 show how the UCM is calculated for each line of code. It also demonstrates how complexity value varies from line to line depending on the architecture and size of the line. The highest complexity value is 30 for line number 31 since this line consists a loop and ten operators and operands. In other words, this line is most complex in its structure and size. On the contrary, complexity value is zero for lines 1, 2, 3, 16, 30, 35 since these lines have the simplest structure, which do not contain any operator or operand. Similarly, line 14 and 16 have function calls and therefore the complexity due to call is double in comparison to an ordinary program line (without any branching, iterations, or embedded systems).

## 4 Experimentation and Comparative Study

Empirical studies play an important role in the evaluation of software engineering discipline [1]. We have taken eight different ‘C’ programs from Misra and Mishra [5] for the analysis of the UCM approach. We calculated the Unified Complexity Measure (UCM) for each one of those programs (see Table-3). The complexity values for their components and UCM are also given in table 3. We observe from this table that the UCM values are high for programs whose program lines generally contain high value for any one of their components. Obviously, it is due the fact that UCM depends on the number of lines, operators, operands and cognitive weights.

We also used these sample programs to calculate the value of four different complexity measures, namely cognitive functional size complexity measure, effort measure, cyclomatic complexity and statement count, for comparative purposes (Table-4). Inspection of Table 4 states that the behavior of UCM is similar to the

**Table 3.** Calculated complexity values for UCM and its Components

| No. | The Number of Lines (NL) | SOO | CW | UCM |
|---|---|---|---|---|
| 1 | 12 | 20 | 4 | 50 |
| 2 | 17 | 35 | 3 | 57 |
| 3 | 18 | 52 | 3 | 71 |
| 4 | 37 | 58 | 16 | 136 |
| 5 | 23 | 25 | 10 | 79 |
| 6 | 15 | 20 | 6 | 57 |
| 7 | 11 | 10 | 6 | 43 |
| 8 | 11 | 17 | 9 | 73 |

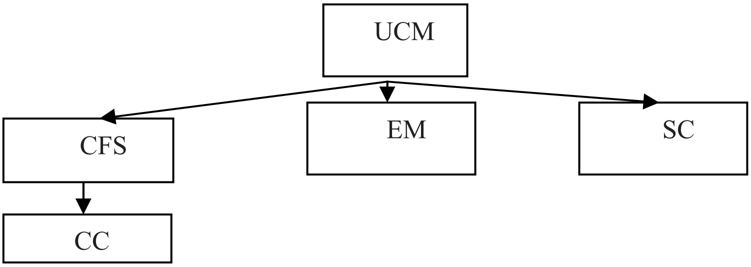


**Fig. 1.** UCM and other related complexity measures. CFS: Cognitive functional Size, EM: Effort Measure, SC: Statement Count; CC: Cyclomatic Complexity.

**Table 4.** Complexity values for different measures

| Complexity Measures | Programs | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Pgm.1 | Pgm.2 | Pgm.3 | Pgm.4 | Pgm.5 | Pgm.6 | Pgm.7 | Pgm. 8 |
| **Statement Count** | 12 | 17 | 18 | 37 | 23 | 15 | 11 | 11 |
| **Cyclomatic Complexity** | 2 | 2 | 2 | 5 | 4 | 2 | 3 | 4 |
| **Effort Measure** | 1859 | 5191 | 6237 | 15556 | 5079 | 2869 | 1221 | 1039 |
| **Cognitive functional size** | 8 | 9 | 9 | 46 | 30 | 14 | 21 | 30 |
| **Unified Complexity Measure** | 50 | 57 | 71 | 136 | 79 | 57 | 43 | 73 |

other complexity measures. The higher values of UCM is due to the fact that the UCM includes most of the parameters of different measures. This means, the UCM can be assumed to be a superset (see fig 1.) of cognitive complexity, effort measure, cyclomatic complexity and statement count measures, which seems to be the most important advantage of UCM.

Interestingly, the inspection of Figure 2 states that the UCM and CFS show almost the same trend but the UCM has higher values. The relatively high values of UCM are because the UCM already includes the considerations of all cognitive aspects of CFS. Especially, the highest value of UCM for the sample program 4 is due to the contribution of other factors i.e. larger size of the code, high cognitive complexity, high occurrences of operators and operands.

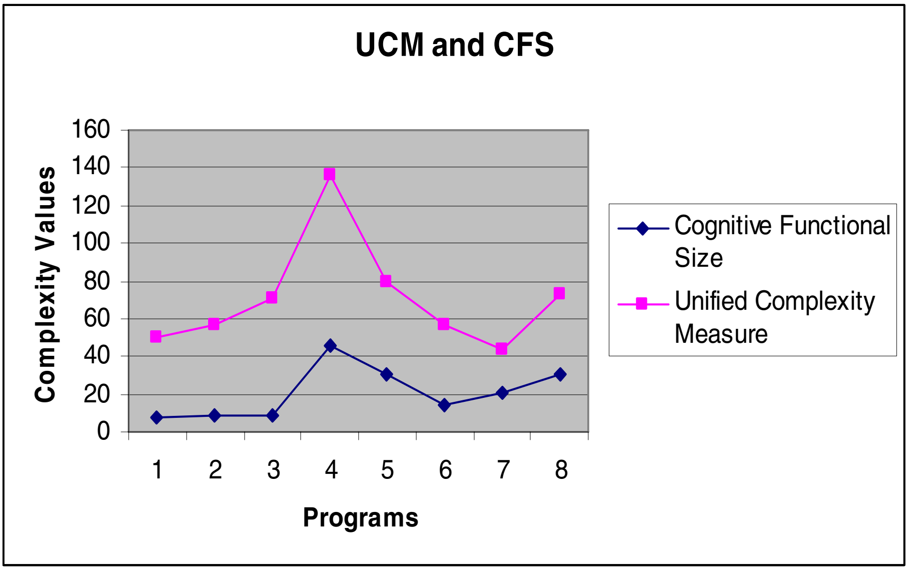


**Fig. 2.** Comparative Graph of UCM with CFS

## 5 Conclusion

In this paper, we proposed a metric by primarily considering all the internal attributes which directly affect the complexity. It uses number of lines (size), total occurrence of operators and operands (size), number of control structures (control flow structuredness) and function calls (coupling) as the internal attributes. The proposed metric also considers cognitive complexity since it is one of the important factors for increasing overall complexity and relates to comprehension. Understandability of software is the program comprehension and is a cognitive process. The cognitive complexity is used in terms of cognitive weights of basic control structures, which is also an indication of structural complexity. This means, the proposed metric is a unique model including all the factors responsible for increasing the complexity. The use of proposed metric is demonstrated by using a simple programming example. The practical applicability of the metric is evaluated by using eight different test cases which prove the soundness and robustness of the proposed measure. As a conclusion, we hope that the proposed metric, UCM, will aid the developers and practitioners in evaluating the complexity before and after coding.

## References

1. Basili, V.: The Role of Controlled Experiments in Software Engineering Research. In: Basili, V.R., Rombach, H.D., Schneider, K., Kitchenham, B., Pfahl, D., Selby, R.W. (eds.) Empirical Software Engineering Issues. LNCS, vol. 4336, pp. 33–37. Springer, Heidelberg (2007)

2. Halstead, M.H.: Elements of Software Science. Elsevier North-Holland, New York (1997)
3. IEEE Computer Society: Standard for Software Quality Metrics Methodology. Revision IEEE Standard, 1061–1998 (1998)
4. McCabe, T.H.: A Complexity Measure. IEEE Transactions Software Engineering, 308–320 (1976)
5. Misra, S., Misra, A.K.: Evaluating Cognitive Complexity Measure with Weyuker's properties. In: Proc. of IEEE (ICCI 2004), pp. 103–108 (2004)
6. Misra, S.: Cognitive Program Complexity Measure. In: Proc. of IEEE (ICCI 2007), pp. 120–125 (2007)
7. Mills, E.: Software Metrics (2007), http://www.sei.UCMu.edu/publications/documents/UCMs/UCM.012.html
8. Oviedo, E.I.: Control flow, Data and Program Complexity. In: Proc. of IEEE COMPSAC, Chicago, IL, pp. 146–152 (1980)
9. Pressman, R.S.: Software Engineering: A Practitioner's approach, 5th edn. McGraw Hill, New York (2001)
10. Sommerville, I.: Software Engineering, 6th edn. Addison-Wesley, Reading (2001)
11. Wang, Y., Shao, J.: A New Measure of Software Complexity based on Cognitive Weights. Can. J. Elect. Comp. Eng. 28(2), 69–74 (2003)
12. Wang, Y.: The theoretical framework of cognitive informatics. International Journal of Cognitive Informatics and Natural Intelligence 1(1), 10–22 (2007)

# Efficient Tracking with AdaBoost and Particle Filter under Complicated Background

Yuji Iwahori[1], Naoki Enda[1], Shinji Fukui[2], Haruki Kawanaka[3], Robert J. Woodham[4], and Yoshinori Adachi[1]

[1] Faculty of Engineering, Chubu University
Matsumoto-cho 1200, Kasugai 487-8501, Japan
iwahori@cs.chubu.ac.jp, enda@cvl.cs.chubu.ac.jp, adachiy@isc.chubu.ac.jp
http://www.cvl.cs.chubu.ac.jp/

[2] Faculty of Education, Aichi University of Education
Hirosawa, Igaya-cho, Kariya 448-8542, Japan
sfukui@auecc.aichi-edu.ac.jp

[3] Faculty of Information Science and Technology, Aichi Prefectural University
Nagakute-cho, Aichi-gun 480-1198, Japan
kawanaka@ist.aichi-pu.ac.jp

[4] Department of Computer Science, University of British Columbia
Vancouver, B.C. Canada V6T 1Z4
woodham@cs.ubc.ca

**Abstract.** Particle filter, which is the probability technique, can be used for the robust tracking to the noise and the occlusion. However, when many objects are tracked simultaneously, the real-time tracking becomes difficult as the computational cost increases. While, the AdaBoost has an ability that it has the remarkable efficiency as a statistical technique in pattern recognition. AdaBoost can be used to detect an object region for the efficient tracking with a particle filter. However, it is difficult to detect the moving object under the complicated background by AdaBoost. This paper proposes an improvement of efficiency of particle filter by introducing further distinction features using AdaBoost for the complicated background.

**Keywords:** Moving Object Tracking, AdaBoost, Particle Filter.

## 1 Introduction

For motion tracking, particle filter [1]-[3] is one of the techniques for robust tracking in the presence of occlusion and noise. Particle filter is called Bayesian filter or Sequential Monte Carlo method (SMC), which are sophisticated model estimation technique based on simulation, and it is used to estimate Bayesian models.

Particle filter uses the posteriori estimation based on the past and the present observations of tracking object. Many particles can archieve the robust tracking but the computational cost increases with the number of particles. Then real-time tracking becomes difficult with the increase of moving objects.

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 887–894, 2008.

In this paper, the effective tracking is proposed without increasing the number of particles. Key idea is that the distributed region of particles would be restricted to realize the effective tracking with the real-time computation. AdaBoost [4] is introduced for the effeciency of particle filter. AdaBoost is available for the problem with the complicated nonlinear classification as a statistical approach of pattern recognition. In recent years, AdaBoost has been introduced for the purpose of face detection [5][6].

This paper proposes a new approach which uses AdaBoost to compensate the computational cost of the particle filter. It shows the combination of the particle filter with AdaBoost makes the performance higher.

It also achieves robust tracking for the case when the observation distribution is non-Gaussian. It approximates the discrete probability density where the random variables are represented by many particles. In this sense, the particle filter is used not only in the field of motion tracking but also in the field of speech recognition or other applications. Some researches treat the motion tracking [7] or human head tracking [8]. Combination of the particle filter with other algorithmsmakes the performance stronger. It is shown that the proposed approach can track the human with AdaBoost under the complicated background.

## 2 Probability Based Tracking

**Time Sequential Filtering.** Time sequential filtering is a method to estimate the most suitable value from the past and present observation values. Let the state of tracking target at time $t$ be $\mathbf{x}_t$, and let the observation result from image be $\mathbf{z}_t$. Let the observation results by time $t$ be $\mathbf{Z}_t = (\mathbf{z}_1, \ldots, \mathbf{z}_t)$. The probability density is discretely approximated by many particles with the state and the likelihood. The robust tracking to both the noise and the variation of environment is performed.

**Weighting Sampling.** Particle filter approximates the posterior $p\left(\mathbf{x}_t | \mathbf{Z}_t\right)$ at time $t$ with $N$ particles which consist of the state $\mathbf{x}$ and its weight. Weight $\pi_t^{(i)}$ for the state $\mathbf{x}_t^{(i)}$ at time $t$ for $i$-th hypothesis is evaluated by the likelihood function $p\left(\mathbf{z}_t | \mathbf{x}_t = \mathbf{x}_t^{(i)}\right)$.

**Particle Filter Based Tracking.** Tracking with hypothesis is realized by repeating the following processes.

1. Sampling of hypotheses $\left\{\mathbf{s}_{t-1}^{\prime(1)}, \ldots, \mathbf{s}_{t-1}^{\prime(N)}\right\}$ using the weight $\pi_{t-1}^{(i)}$ based on the state $\mathbf{x}_{t-1}^{(i)}$ of particles $\left\{\left(\mathbf{x}_{t-1}^{(i)}, \pi_{t-1}^{(i)}\right), i = 1, \ldots, N\right\}$ which approximates the posterior distribution $p\left(\mathbf{x}_{t-1} | \mathbf{Z}_{t-1}\right)$ at time $t-1$.
2. Generate $N$ hypotheses $\mathbf{x}_t^{(i)}$ at time $t$ from the sampled hypotheses $\mathbf{s}_{t-1}^{\prime(i)}$.
3. Likelihood function from $\mathbf{x}_t^{(i)}$ and the weight $\pi_t^{(i)}$ of $\mathbf{x}_t^{(i)}$ are calculated. Here, the weight is normalized so that $\sum_{i=1}^{N} \pi_t^{(i)} = 1$ holds.

Particles $\left\{\left(\mathbf{x}_t^{(i)}, \pi_t^{(i)}\right), i = 1, \ldots, N\right\}$ are obtained as a discrete approximation of posterior distribution $p\left(\mathbf{x}_t|\mathbf{Z}_t\right)$ at time $t$. Mean value of the hypotheses is used as the estimated state for the tracked target at time $t$.

## 3 AdaBoost

AdaBoost is one of the boosting techniques among ensemble learning. Weight of learning data is updated for weak learning unit with weak classification in AdaBoost. The decision of majority based on the weight of weak learning unit determines the final result. AdaBoost was proposed by Freund *et al.* [4] to solve the early boosting problems, and it gives great performance to pattern recognition. This method uses the AdaBoost for classification of binary distinction.

A set of weak learning units $f_j(x) : j \in 1, \cdots, J$ is prepared to take the value of ±1 for the example data set $(x_i, y_i) : i = 1, \cdots, n$. Here, $n$ represents the number of examples and $j$ represents the number of weak learning units. The algorithm of AdaBoost is represented as follows.

```
Learning Algorithm of AdaBoost
(1) Initialization of Weights
      $\omega_1(i) = \frac{1}{n}, (i = 1, \cdots, n)$
(2) Boosting Round Error rate with weight at round $t = 1, \cdots, T$
      $\epsilon_t(f) = \sum_{i=1}^{n} I(f(x_i) \neq y_i)\omega_t(i)$
      $I$ is set to be 1 when the output of weak learning unit and given example becomes different, else $I$ is set to be 1.
    (2.1) Selection of weak learning unit with minimum error rate
          $f_t = \text{arg min } \epsilon_t(f_j)$
    (2.2) (Union Weight)
          $\beta_t = \frac{1}{2} \log \frac{1 - \epsilon_t(f_t)}{\epsilon(f_t)}$
    (2.3) Updating weight
          $\omega_{t+1}(i) = \frac{\omega_t(i) \exp\{-y_i \beta_t f_t(x_i)\}}{\sum_{k=1}^{n} \omega_t(k) \exp\{-y_k \beta_t f_t(x_k)\}}$
(3) Decision by majority with weight.
      $f = \text{sign}(F_t(x)), \;\; F_t = \sum_{t=1}^{T} \beta_t f_t(x)$
```

AdaBoost constructs the efficient combination of weak learning units with updating weight. The weak learning unit searched for next round can compensate the previous weak weight unit. Thus, AdaBoost does not require the specific parameters and it is a good candidate to high ability classification.

# 4 Efficiency of Particle Filter with AdaBoost

Particle filter generates the particles at time $t$ with random variables based on the estimated status obtained from sampling at time $t-1$. Instead of distributing particles to wider region except the target object, particle filter increases the processing of calculating the likelihood function using the color histogram [9] inside the rectangular region. Robust tracking needs many particles to track the object but this processing cost becomes sometimes problem with increasing the target objects.

In this paper, AdaBoost is introduced to detect the moving object and restricts the rectangular region of distribution of particles. AdaBoost detects the distributed region with simple calculation and fast processing, then particle filter estimates the tracking object inside the detected region with color histogram. This processing makes it possible to reduce the number of particles used, and the total processing time can be saved even through AdaBoost increases the processing time for the detection of moving object.

## 4.1 Distinction Features

This approach uses the distinction feature to obtain the tracking object region in the image with small cost. The distinction feature is used to AdaBoost to distinguish the region whether it is object region or not with fast processing. A total of 8 features is introduced to perform the robust detection of moving object region. First features $D_R, D_G, D_B$, are used as the background subtraction of RGB color values. Background subtraction is taken for each color $c \in \{R, G, B\}$ and the mean value in the window size of $X \times Y$ is adopted as the feature.

The second feature $D_f$ is the difference between frames. The difference between frames is taken for monochrome image converted from color image. The mean value of the difference between frames for the window size is used as the feature.

Further features are edge background subtraction. Edge detection is done for each direction $v \in \{V, H, R, L\}$. The mean values for these four directions based on the following edge operators are taken as 4 features.

$$h_V = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \quad h_H = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}$$

$$h_R = \begin{bmatrix} 0 & 1 & 2 \\ -1 & 0 & 1 \\ -2 & -1 & 0 \end{bmatrix} \quad h_L = \begin{bmatrix} 2 & 1 & 0 \\ 1 & 0 & -1 \\ 0 & -1 & -2 \end{bmatrix}$$

To reduce the calculation cost and time, lower resolution processing for the distinction region from $X \times Y$ pixels to $\frac{X}{5} \times \frac{Y}{5}$ pixels is used for AdaBoost. Let the input image $G(x, y)$, the edge strength $e_v$ for the direction $v$ is

$$e_v(x, y) = \left| \sum_{m=0}^{2} \sum_{n=0}^{2} G(x+m, y+n) h_v(m, n) \right| \tag{1}$$

The edge subtraction feature $E_v$ is given as

$$E_v = \frac{\sum_{i=0}^{(X-1)/5} \sum_{j=0}^{(Y-1)/5} |e_v^n(x+i, y+j) - e_v^b(x+i, y+j)|}{\frac{X \times Y}{25}} \tag{2}$$

A total of 8 dimensional feature vector $\boldsymbol{V}$ is used in AdaBoost.

$$\boldsymbol{V} = [D_R \;\; D_G \;\; D_B \;\; D_f \;\; E_V \;\; E_H \;\; E_R \;\; E_L]^T \tag{3}$$

$\boldsymbol{V}$ is used to detect where the window region is in the tracking region or not. Learning data is given as ±1 for each answer of features. This approach takes +1 for the tracking target region, while −1 for the background region.

When the feature vector consists two dimensions, the kind of distinction of weak learning unit becomes four. Therefore, eight dimensional feature set produces sixteen kinds of distinction of weak learning units.

Weighted learning data is distinguished into sixteen classes. The result of distinction with least error for the various threshold value is given as the distinction of Boosting Round $t$. Connected weight $\beta_t$ becomes the precision based on the error at the distinction. Here, the threshold operation is iterated for the weighted learning data at the Boosting Round. This simple threshold operation as shown in Fig. 1 works as the complicated function at the weighted decision by majority. Boosting ends when the connected weight becomes less than 1% of the initial value.

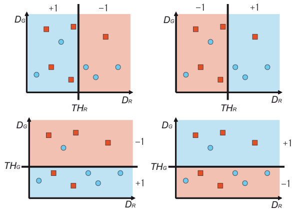


**Fig. 1.** Kinds of Distinction

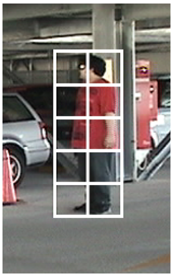


**Fig. 2.** Object Region by AdaBoost

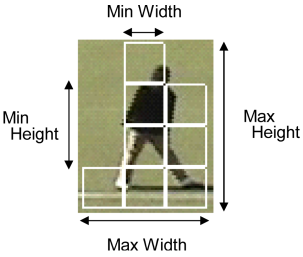


**Fig. 3.** Maximum and Minimum Values for Width and Height of Rectangle

### 4.2 Restriction of Distribution of Particles

Windows of $X \times Y$ pixels are aligned without overlapping for the region of distribution of particles based on the center location of the tracking result obtained from the previous frame. Eight dimensional feature $\boldsymbol{V}$ is obtained from each of the aligned windows. AdaBoost learning distinction function estimates whether each window belongs to the tracking region or not. Adding result of whole weighted output of weak learning units estimates the tracking region when the output of strong learning unit $f$ becomes $f \geqq 0$ while it estimates the background region when $f < 0$. Fig. 2 shows the distinction result to unknown data. White rectangle shows the region of tracking object by AdaBoost. This result is used to restrict the distribution of each particle inside the region of tracking object.

From the result of AdaBoost for unknown data at time $t$, the maximum value and the minimum value for the rectangular width and height is obtained to distribute particles. Fig. 3 shows the maximum and the minimum values for the rectangle recognized as the tracking target region for each target. Let the number of horizontal blocks be $m$, number of horizontal pixels of a block be $p$, then the maximum value is set to be $m \times p$, and the minimum value is set to be $(m-2) \times p$ when $m > 2$. While the maximum value is set to be $m \times p$, and the minimum value is set to be $(m-1) \times p$ when $n \leqq 2$. The maximum and minimum value for the height of rectangle is obtained in the same way.

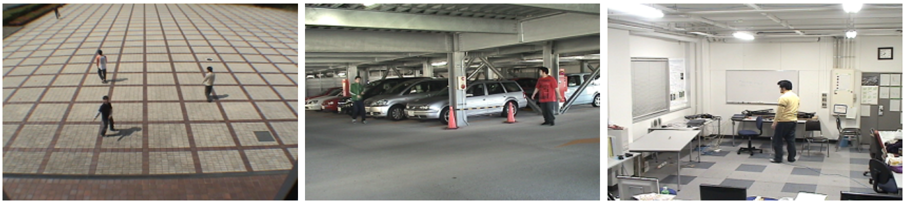
## 5 Experiments

### 5.1 Learning and Detection of Tracking Object Using AdaBoost

$720 \times 480$ pixel image sequence is used with the Core 2 Duo E6600 with Main Memory 1024 MBytes in the experiments. 20,000 example data were used to construct the distinction function as the learning data set. 10,000 examples were data in the tracking region, while 10,000 examples were not in the tracking region. The constructed AdaBoost was used to detect the tracking region with the distinction function based on the weighted decision by majority for unknown data using combination of weak learning unit.

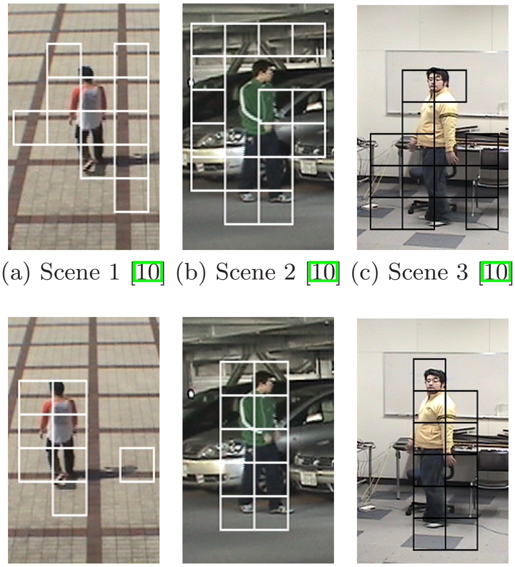
The proposed approach extends the method [10] to the tracking of moving object under more complicated background. The performance of distinction was evaluated in comparison with [10] and this extended approach.

50 frames in video sequence shown in Fig. 4 were used in the experiments. The size of window region was $30 \times 30$ pixels for Scene 1 and 2, and $40 \times 40$ pixels for Scene 3. As the comparison of [10] with the proposed approaches, detection



(a) Scene 1 (b) Scene 2 (c) Scene 3

**Fig. 4.** Scene in Image Sequence



(a) Scene 1 [10] (b) Scene 2 [10] (c) Scene 3 [10]

(a) Scene 1 (Proposed) (b) Scene 2 (Proposed) (c) Scene 3 (Proposed)

**Fig. 5.** Detection Results

ratio, non-detection ratio, and wrong detection ratio were evaluated. Examples of result detected by AdaBoost are shown in Fig. 5, respectively. Manual checking with eyes was used for the evaluations. It was confirmed that [10] has 81 to 83 % detection ratio while proposed approach has 98 %. The non-detection ratio was 16 to 29 % in [10] while 1.2 to 1.76 % in this approach. The wrong detection ratio was 31 to 44 % in [10] while 5 to 10 % in this approach. The better results were obtained under complicated background. Some detection errors are caused from that shadow regions or the moving object is mirrored to the cars.

### 5.2 Efficiency of Particle Filter

We further confirmed the efficiency of particle filter with AdaBoost detection. Particle filter without AdaBoost detection needs at least 30 particles for each object to realize the robust tracking. When AdaBoost was further introduced, 5 particles could still realize the robust tracking. When number of particles becomes smaller, the proposed approach can continue more robust tracking than [10]. The experiment was done for the video sequence including the tracking object with 145 × 60 pixels. AdaBoost with the particle filter could realize the fast and good performance through a total processing. The processing time could be saved with introducing AdaBoost in spite of adding the distinction processing with AdaBoost.

## 6 Conclusion

This paper proposed a new method to improve the efficiency of particle filter based tracking with AdaBoost for the complicated background. Feature vector which is robust to the complicated background was considered and introduced to AdaBoost distinction function, while [10] was able to be applied to the tracking object under the simple background. The approach can restrict the distribution of particles to the tracking object region by AdaBoost detection under the complicated background. Smaller number of particles are enough to continue the stable tracking. As a result, the total processing time could be reduced by the efficient processing with AdaBoost and particle filter.

Further subjects are remained for the cases when shadow region exists or the tracking object is mirrored to another object in the image. Dynamic determination of window region with AdaBoost distinction is also a remained problem.

## Acknowledgment

Iwahori's research is supported by JSPS Grant-in-Aid for Scientific Research (C)(20500168), Chubu University Grant, Nitto Foundation. Wooodham's research is supported by the Natural Sciences and Engineering Research Council (NSERC).

## References

1. Isard, M., Blake, A.: CONDENSATION - conditional density propagation for visual tracking. Intl. J. of Computer Vision 29(1), 5–28 (1998)
2. Doucet, A., Godsill, S., Andrieu, C.: On sequential Monte Carlo sampling methods for Bayesian filtering. Statistics and Computing 10(3), 197–208 (2000)
3. Liu, J.S., Chen, R.: Sequential Monte Carlo Methods for Dynamic Systems. Journal of the American Statistical Association 93(443), 1032–1044 (1998)
4. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. J. of Computer and System Sciences 55(1), 119–139 (1997)
5. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proc. of CVPR, vol. 1, pp. 511–518 (2001)
6. Huang, C., Wu, B., Ai, H., Lao, S.: Omni-directional face detection based on real AdaBoost. In: ICIP, vol. 1, pp. 24–27 (2004)
7. Iwahori, Y., Kawanaka, H., Takai, T., Adachi, Y., Itoh, H.: Particle Filter Based Tracking of Moving Object from Image Sequence. In: Gabrys, B., Howlett, R.J., Jain, L.C. (eds.) KES 2006. LNCS (LNAI), vol. 4252, pp. 401–408. Springer, Heidelberg (2006)
8. Kawanaka, H., Fujiyoshi, H., Iwahori, Y.: Human Head Tracking in Three Dimensional Voxel Space. Proceedings of ICPR 3, 826–829 (2006)
9. Swain, M.J., Ballard, D.H.: Color Indexing. Intl. J. Computer Vision 7(1), 11–32 (1991)
10. Kawanaka, H., Ohnishi, K., Iwahori, Y., Fukui, S.: Efficiency of Particle Filter Using Adaboost, Symposium of Meeting on Image Recognition and Understanding. In: Proceedings of MIRU 2006, pp. IS1–5 (2006)

# Analog VLSI Layout Design and the Circuit Board Manufacturing of Advanced Image Processing for Artificial Vision Model

Masashi Kawaguchi[1], Takashi Jimbo[2], and Naohiro Ishii[3]

[1] Department of Electrical & Electronic Engineering, Suzuka National College of Technology, Shiroko, Suzuka Mie 510-0294, Japan
masashi@elec.suzuka-ct.ac.jp
[2] Department of Environmental Technology and Urban Planning Graduate School of Engineering, Nagoya Institute of Technology,
Gokiso-cho, Showa-ku, Nagoya, 466-8555 Japan
jimbo.takashi@nitech.ac.jp
[3] Department of Information Science, Aichi Institute of Technology,
Yachigusa, Yagusa-cho, Toyota, 470-0392 Japan
ishii@aitech.ac.jp

**Abstract.** We propose herein an artificial vision model for the motion detection which uses analog electronic circuits and design the analog VLSI layout. We also have manufactured the circuit board. The proposed model is comprised of four layers. The model was shown to be capable of detecting a movement object. The number of elements in the model, is reduced in its' realization using the integrated devices. The proposed model is robust with respect to fault tolerance. Moreover, the connection of this model is between adjacent elements, making hardware implementation easy.

**Keywords:** Neural Network, Motion Detection, Analog Circuits, Biomedical Vision System.

## 1 Introduction

A neuro chip and an artificial retina chip are developed to comprise the neural network model and simulate the biomedical vision system. At present, a basic image processing, such as edge detection and reverse display of an image has been developed [1][2]. The retina consists of the inside retina and outside retina. The inside retina sends the nerve impulses to the brain, whereas the outside retina receives optical input from the visual cell. As a result, the outside retina emphasizes spatial changes in optical strength. Recently, the network among the amacrine cell, the bipolar cell and the ganglion cell has been clarified theoretically, which has led to active research concerning the neuro-device, which models the structure and function of the retina. Easy image processing, reversing, edge detection, and feature detection, have been achieved by technologies such as the neuro chip and the analog VLSI circuit.

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 895–902, 2008.

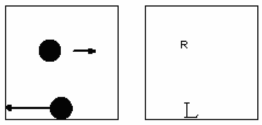


**Fig. 1.** Example of Advanced Image Processing

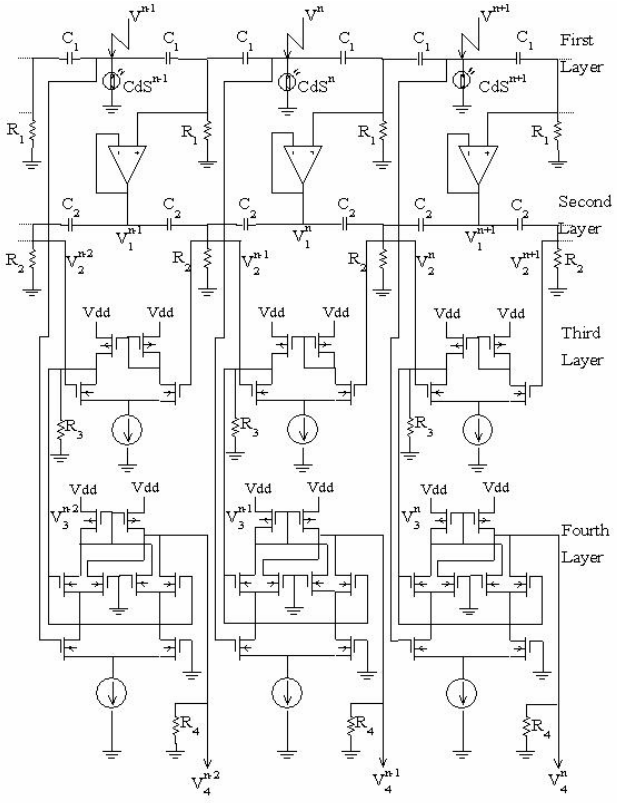


**Fig. 2.** One-Dimensional Four-Layered Direction Model for Selective Motion Detection

Some motion detection models are proposed in the recent researches. Figure 1 shows the example of advanced image processing. It is direction sensitive motion detection behavior. When the object moves from left to right slowly the model outputs a small "right" signal, and when the object moves from right to left quickly the model outputs a big "left" signal.

## 2 Advanced Image Processing

Lu et al. describes the application of an analog VLSI vision sensor to active binocular tracking. The sensor outputs are used to control the vergence angles of the two cameras and the tilt angle of the head so that the center pixels of the sensor arrays image the same point in the environment[3]. Another model presents the implementation of a visual motion detection algorithm on an analog network. The algorithm in the model is based on Markov random field (MRF) modeling. Robust motion detection is achieved by using a spatiotemporal neighborhood for modeling pixel interactions. Not only are the moving edges detected, but also the inner part of moving regions [4]. The other model is an analog MOS circuit inspired by an inner retina. The analog circuit produces signals of motion of edges which are output in an outer retinal neural network. Edge signals are formed into half-wave rectified impulses in two types of amacrine cells, and fed back to the wide field amacrine cell in order to modulate width of impulses [5]. In the present study, we propose a motion detection model in which the speed is detected by differentiation circuits. The surface layer is composed of the connections of capacitors. In the inner layer, the movement direction is detected by difference circuits. When the object moves from left to right, a positive output signal is generated, and when the object moves from right to left, a negative output signal is generated. We show this model is able to detect the speed and direction of a movement object by the simple circuits. Despite the large object size, this model can detect the motion.

## 3 One Dimensional Motion Detection Model

We first developed a one-dimensional model, the structure of which is shown in Fig. 2.

### 3.1 First Layer Differentiation Circuits (First Layer)

The current is given by equation (1), where the input voltage is denoted by $V^n$ and the capacitance is denoted by $C_1$. The current into a capacitor is the derivative with respect to time of the voltage across the capacitor, multiplied by the capacitance.

$$I = C_1 \frac{dV^n}{dt} \tag{1}$$

$$V_1^n = IR_1 = C_1 R_1 \frac{dV^n}{dt} \tag{2}$$

The output voltage $V_1^n$ is given by equation (2). Equation (2) is multiplied by the resistance $R_1$, calculating the voltage potential. Buffer circuits are realized by operational amplifiers between the first layer and the second layer. In the first layer, there are also the CdS Photoconductive Cells. Using CdS cells, this model is not affected by object luminance. When the object is high luminance, the resistances of CdS cells are low. Some currents flows to ground through the CdS. Therefore, despite the high luminance, the input Voltage $V_1^n$ is not affected.

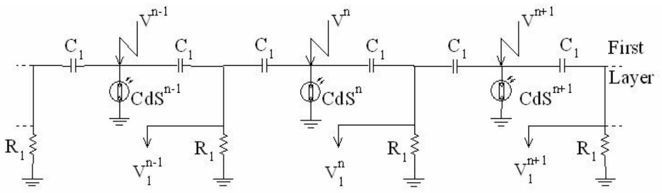


**Fig. 3.** First Layer Differentiation Circuits

### 3.2 Second Layer Differentiation Circuits (Second Layer)

The second Layer is also composed of differentiation circuits; however, the CR coefficient is small compared that of the first layer differentiation circuits. The output of first layer, $V_1{}^{\mathrm{n}}$, is differentiated again, and the output of the second layer is assumed to be $V_2{}^{\mathrm{n}}$, calculating the voltage potential.

$$I = C_2 \frac{dV_1^n}{dt} \tag{3}$$

$$V_2^n = IR_2 = C_2 R_2 \frac{dV_1^n}{dt} \tag{4}$$

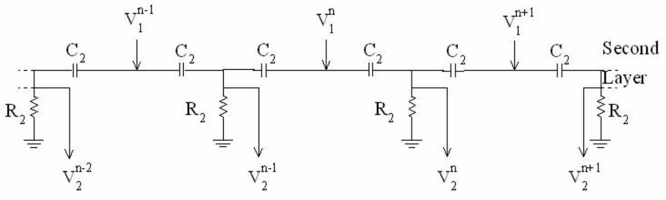


**Fig. 4.** Second Layer Differentiation Circuits

### 3.3 Difference Circuits (Third Layer)

The third layer consists of difference circuits realized by MOSFET. The bottom $I_b$ is a current source. The manner in which $I_b$ is divided between $Q_1$ and $Q_2$ is a sensitive function of the difference between $V_2{}^{\mathrm{n+1}}$ and $V_2{}^{\mathrm{n}}$, and is the essence of the operation of the stage. We assume the MOSFET device is in the sub-threshold region and the *I-V* characteristics follows the exponential characteristics, then the drain current $I_D$ in the sub-threshold region is exponential in the gate voltage $V_g$ and source voltage $V_s$. $V$ is electric potential of current source $I_b$. $I_0$ and κ are coefficients.

The circuit consists of a differential pair and a single current mirror, like the one shown in Figure 2, which is used to subtract the drain currents $I_1$ and $I_2$. The current $I_1$ drawn out of $Q_3$ is reflected as an equal current out of $Q_4$; the output current $I_{out}$ is thus equal to $I_1$-$I_2$, and is therefore given by Equation (5).

The output voltage of this circuit is as follows.

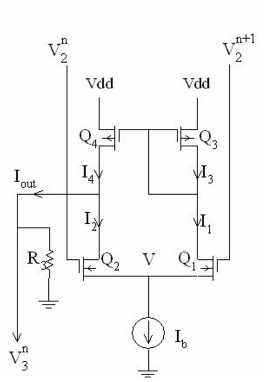


**Fig. 5.** Difference circuits(Third layer)

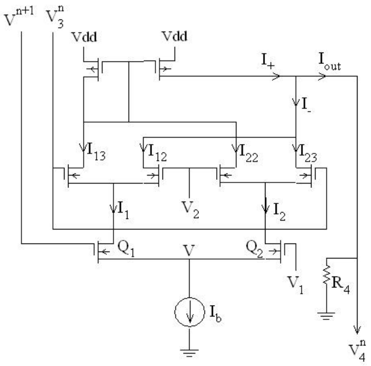


**Fig. 6.** Gilbert multiple circuits (fourth layer)

$$\begin{aligned} I_1 - I_2 &= I_b \frac{e^{\kappa\left(V_2^{n+1}-V_2^n\right)/2} - e^{-\kappa\left(V_2^{n+1}-V_2^n\right)/2}}{e^{\kappa\left(V_2^{n+1}-V_2^n\right)/2} + e^{-\kappa\left(V_2^{n+1}-V_2^n\right)/2}} \\ &= I_b \tanh \frac{\kappa(V_2^{n+1} - V_2^n)}{2} \end{aligned} \tag{5}$$

$$V_3^n = (I_1 - I_2)R_3 = I_b R_3 \tanh \frac{\kappa(V_2^{n+1} - V_2^n)}{2} \tag{6}$$

### 3.4 Gilbert Multiple Circuits (Fourth Layer)

The fourth layer is comprised of Gilbert multiple circuits shown in Figure 6. We assume the MOSFET device is in the sub-threshold region, the *I*-*V* characteristics follows the exponential characteristics, then the drain current $I_D$ in the sub-threshold region is exponential in the gate voltage $V_g$ and source voltage $V_s$. The result for the two drain currents of the differential pair were derived. In Figure 1, $V_1$ and $V_2$ are connected to ground respectively. In this circuit, the voltage $V_1$ and $V_2$ are 0. The fourth layer produces the third layer output $V_3^n$ and the input signal $V^{n+1}$. This circuit detects the pure output of movement. $I_b$ is the current source, and κ is a coefficient.[1]

$$I_4^n = I_b \tanh \frac{\kappa V_3^n}{2} \tanh \frac{\kappa V^{n+1}}{2} \tag{7}$$

$$V_4^n = I_4^n R_4 = I_b R_4 \tanh \frac{\kappa V_3^n}{2} \tanh \frac{\kappa V^{n+1}}{2} \tag{8}$$

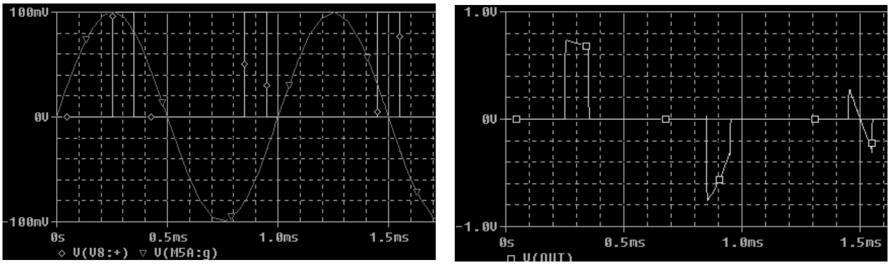


**Fig. 7.** Experimental results of Gilbert multiple circuits

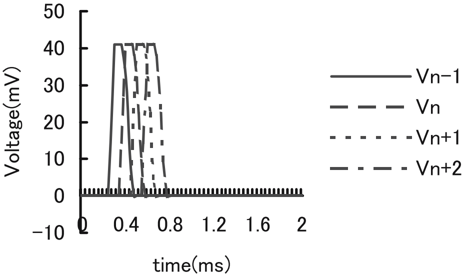


**Fig. 8.** Output of the fourth layer

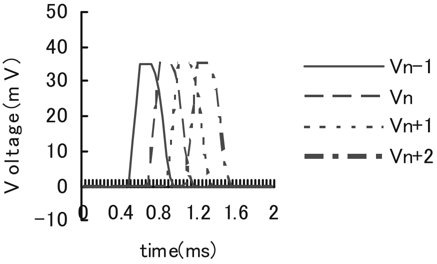


**Fig. 9.** Output of the half speed of input (after multiple circuit processing)

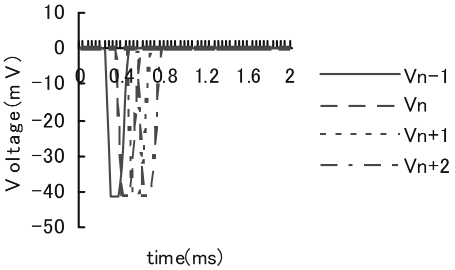


**Fig. 10.** Output of the reverse input direction

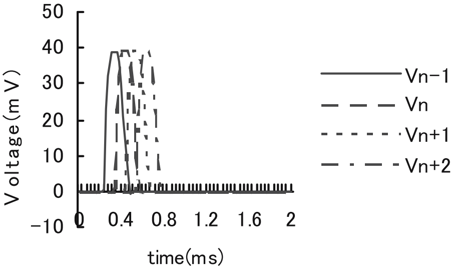


**Fig. 11.** Output when the object size is large

$I_4^n$ is the output current of the fourth layer, $R_4$ is the earth resistance, and $V_4^n$ is the final output. $I_4^n$ corresponds to $I_{out}$ in Figure 6. Using multiple circuits, this model can detects the pure output of movement. We set the parameter of circuits as follows. In the first layer, $C_1$=0.1μF, $R_1$=1kΩ. We used the μA741 as a buffer circuits. In the second layer, $C_2$=0.1μF, $R_2$=100kΩ. At the difference circuits, we used the VP1310 and VN1310 as MOSFET [6]. The connection of this model is between adjacent elements, making hardware implementation easy. We measured the shape of the output waves produced by the input movement signal using an electronic circuit simulator

(SPICE). Figure 8 shows the final output, which indicates that this circuit detects the pure output of the movement. Figure 9 shows the final output when the object moves at half speed. Figure 10 shows the final output when the object moves from right to left. When the object moves from right to left, this model outputs a negative signal. Figure 11 shows the final output when the object size is large, which is nearly identical to the result shown in Fig. 8, indicating that this model is not affected by object size, with respect to speed detection. These results shows that this model is able to detect the speed and direction of a movement object in one dimension.

### 3.5 Designs for Circuit Board

Next, we designed the circuit board using CAD system by MITS Corporation. This data is for making the circuit board using manufacturing system. In this paper, we show that it is realized this model by the real circuit, not simulation.

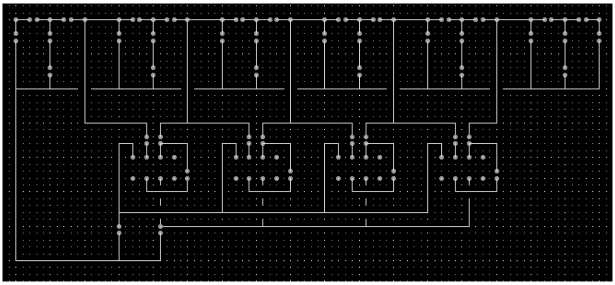


**Fig. 12.** Circuit Board Manufacturing of One-Dimensional Model

### 3.6 Layout for Motion Detection Circuits

The proposed model is processed by the analog electronic circuits. We designed the simulated circuit to the chip layout using Orcad Layout Tool. We show that it is possible to realize the hardware implementation. Figure 13 shows the layout of analog circuit of the one-dimensional model. In the biomedical brain, information is also processed in an analog manner. In the future, movement information will be collected into an analog electronic brain model. This would allow the hardware system of the biomedical brain model to be realized. The proposed moving detection model has possible application as a sensor and can compose part of the receptor. The proposed model will enable the clarification of the mechanism of the biomedical brain.

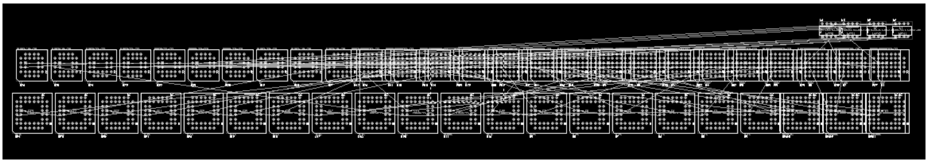


**Fig. 13.** Layout of Analog Circuit of One-Dimensional Model

## 4 Conclusion

We designed the motion detection analogue electric circuit using a biomedical vision system. We first designed the one-dimensional model and experimented. Using the one-dimension model, the movement information was detected. The input terminal and the output terminal were arranged in an alternating manner. As a result, a simple circuit and an equivalent output result were obtained. The realization of an integration device will enable the number of elements to be reduced. The proposed model is robust with respect to fault tolerance. Moreover, the connection of this model is between adjacent elements, making hardware implementation easy. Finally, we designed the layout of analog VLSI model. We show that its model is possible to realize the hardware implementation[7].

**Acknowledgments.** This research was financially supported by Grant-in-Aid from the Japanese Ministry of Education, Science, Sports and Culture.

## References

1. Mead, C.: Analog VLSI and Neural Systems. Addison Wesley Publishing Company, Inc., Reading (1989)
2. Chong, C.P., Salama, C.A.T., Smith, K.C.: Image-Motion Detection Using Analog VLSI. IEEE Journal of Solid-State Circuits 27(1), 93–96 (1992)
3. Lu, Z., Shi, B.E.: Subpixel Resolution Binocular Visual Tracking Using Analog VLSI Vision Sensors. IEEE Transactions on Circuits and Systems-II: Analog and Digital Signal Processing 47(12), 1468–1475 (2000)
4. Luthon, F., Dragomirescu, D.: A Cellular Analog Network for MRF-Based Video Motion Detection. IEEE Transactions on Circuits and Systems-I: Fundamental Theory and Applications 46(2), 281–293 (1999)
5. Yamada, H., Miyashita, T., Ohtani, M., Yonezu, H.: An Analog MOS Circuit Inspired by an Inner Retina for Producing Signals of Moving Edges, Technical Report of IEICE, NC99-112, pp.149–155 (2000)
6. Kawaguchi, M., Jimbo, T., Umeno, M.: Motion Detecting Artificial Retina Model by Two-Dimensional Multi-Layered Analog Electronic Circuits. IEICE Transactions, E86-A-2, pp. 387–395 (2003)
7. Kawaguchi, M., Jimbo, T., Umeno, M.: Analog VLSI Layout Design of Advanced Image Processing For Artificial Vision Model. In: IEEE International Symposium on Industrial Electronics, ISIE 2005 Proceeding, vol. 3, pp. 1239–1244 (2005)

# Classification of Local Surface Using Neural Network and Object Rotation of Two Degrees of Freedom

Takashi Kojima[1], Yuji Iwahori[2], Tsuyoshi Nakamura[1], Shinji Fukui[3],
Robert J. Woodham[4], and Hidenori Itoh[1]

[1] Department of Computer Science and Engineering, Nagoya Insititute of Technology
Gokiso-cho, Showa-ku, Nagoya 466-8555, Japan
kojima@juno.ics.nitech.ac.jp, tnaka@nitech.ac.jp, itoh@nitech.ac.jp
[2] Faculty of Engineering, Chubu University
Matsumoto-cho 1200, Kasugai 487-8501, Japan
iwahori@cs.chubu.ac.jp
http://www.cvl.cs.chubu.ac.jp/
[3] Faculty of Education, Aichi University of Education
Hirosawa, Igaya-cho, Kariya 448-8542, Japan
sfukui@auecc.aichi-edu.ac.jp
[4] Department of Computer Science, University of British Columbia
Vancouver, B.C. Canada V6T 1Z4
woodham@cs.ubc.ca

**Abstract.** Gaussian curvature encodes important information about object shape. This paper presents a technique to classify a local surface into several classes from multiple images acquired under different conditions of illumination. Previous approaches require a separate calibration sphere as a reference object, while the proposed approach requires no calibration object like a sphere. Instead, a target object is rotated with some fixed angles in both the vertical and the horizontal directions and the target object itself generates a virtual sphere. In our recent work, only the geometrical calculation is employed to generate a virtual sphere, however this geometrical calculation causes the error between actual marker position and estimated position based on the assumption of the orthographic projection. To generate the virtual sphere with higher accuracy, we adopt a neural network approximation, which is introduced to achieve high accuracy of the virtual sphere image. Experiments with real data are demonstrated.

**Keywords:** Gaussian Curvature, Surface Classification, Neural Network, Self-Calibration, Photometric Stereo.

## 1 Introduction

Gaussian curvature is a representation to describe the structure of object surface at 3D models. Gaussian curvature can reflect the shape variation of the object, and describe and classify surfaces. The value of Gaussian curvature of the surface is also viewpoint invariant, and it is used in the pattern recognition, object

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 903–910, 2008.

recognition and shape recovery. In the previous approaches, Woodham [1] proposed a method to recover local surface orientation and surface curvature using photometric stereo. Photometric stereo uses multiple shading images of a test object taken from a fixed viewpoint under different light source directions. Iwahori and Woodham [2] [3] et al. have pursued neural network implementations of photometric stereo. Neural network replaces LUT (lookup table) as a way to do non-parametric functional approximation. Angelopoulou and Wolf [4], Okatani and Deguchi [5] have recovered the sign of Gaussian curvature from three images taken under the different light source directions without knowing the values of the surface gradient or correct light source directions. These methods are limited to the diffuse reflectance. Iwahori et al. proposed a method to classify the local surface curvature using RBF neural network [6] for the non-diffuse surface, then extended a method to recover the relative magnitude of Gaussian curvature from shading images [7]. The previous approaches used a calibration sphere with the same reflectance property as a test object [1][2][3] [6][7]. A calibration sphere can be used as a learning object for neural network [2][3] [6][7]. Hertzmann and Seitz [8] proposed a method to compute the geometry of objects. The approach removed the restriction that the reference object must be composed of exactly the same material as the target object. It is assumed that one or more example objects with similar materials and known geometry are imaged under the same illumination conditions.

In this paper, we propose a method to classify a local surface from the target object itself without using a calibration sphere. Instead, the approach uses self-calibration of 2 DOF (Degrees of Freedom) to remove the limitations. The target object itself is rotated in the vertical and horizontal directions. In our recent work[9], a geometrical calculation is employed to assume the position of the marker, however this geometrical calculation will occur the error between actual marker position and assumed position based on the assumption of the orthographic projection. When the position error is large, the classification of the local surface takes mistake. The neural network approach attempts to improve the marker position. Experiments with the self-calibration are shown to classify the local surface.

## 2 Photometric Stereo

### 2.1 Empirical Constraint

The principle of photometric stereo uses three light source to determine the surface normal vector $(n_1, n_2, n_3)$ from the observed image irradiances $(E_1, E_2, E_3)$ locally. While, this approach tries to get the local curvature information directly from $(E_1, E_2, E_3, E_4)$ at the local four points on the test object with non-uniform albedo.

Let the image irradiances at $(x_{obj}, y_{obj})$ on the test object be $(E_{1obj}, E_{2obj}, E_{3obj}, E_{4obj})$ and those at $(x_{sph}, y_{sph})$ on the sphere be $(E_{1sph}, E_{2sph}, E_{3sph}, E_{4sph})$.

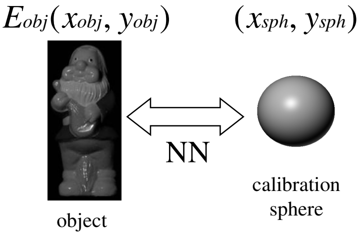


**Fig. 1.** Photometric stereo based on neural network

**Table 1.** Relation between local surface classes and principal curvature

| | $k_2 > 0$ | $k_2 = 0$ | $k_2 < 0$ |
|---|---|---|---|
| $k_1 > 0$ | convex | convex parabolic | hyperbolic |
| $k_1 = 0$ | — | plane | concave parabolic |
| $k_1 < 0$ | — | — | concave |

If the surface material is the same for both of a test object and a sphere, the following constraint

$$\begin{cases} E_{1obj}(x_{obj}, y_{obj}) = E_{1sph}(x_{sph}, y_{sph}) \\ E_{2obj}(x_{obj}, y_{obj}) = E_{2sph}(x_{sph}, y_{sph}) \\ E_{3obj}(x_{obj}, y_{obj}) = E_{3sph}(x_{sph}, y_{sph}) \\ E_{4obj}(x_{obj}, y_{obj}) = E_{4sph}(x_{sph}, y_{sph}) \end{cases} \tag{1}$$

is satisfied. This constraint for $E_1$ to $E_4$ leads to the condition that the corresponding surface normal vector should be the same between a test object and a sphere. This constraint is used to classify the local surface curvature of a test object using neural network [6].

### 2.2 Local Surface Classes

Let the maximum curvature be $k_1$ and let the minimum curvature be $k_2$. Then, the signs of the two principal curvatures $k_1$ and $k_2$ classify local surface into six classes (Table 1). An RBF neural network realizes non-parametric functional approximation in multidimensional spaces. Here, an RBF neural network learns the mapping of ($E_1$,$E_2$,$E_3$,$E_4$) to ($x_{sph}$,$y_{sph}$) for each point on a calibration sphere. The resulting network generalizes in that it predicts a point ($x_{sph}$,$y_{sph}$) on the sphere, given any input values ($E_1$,$E_2$,$E_3$,$E_4$) on the test object. Fig.1 illustarates the overview mentioned above. When we take the local five points of a test object, we can investigate how these points are mapped onto a sphere. The ordering of these mapped points ($x_{sph}$,$y_{sph}$) on a calibration sphere can be used to recover the sign of Gaussian curvature and to classify the local surface into the six classes given in Table 1. The local surface class can be determined from the coordinate of a sphere as the output of neural network.

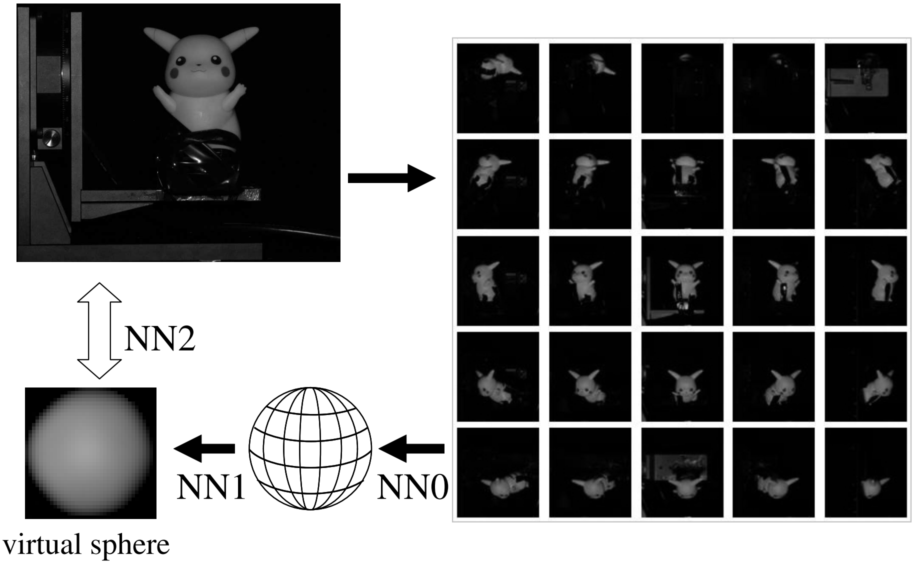


**Fig. 2.** Two DOF Self-Calibration

## 3 Self-calibration and Neural Network for Estimating Marker Position

The novel idea in this paper is to obtain the training data for a neural network (NN2 in Fig.2) from the target object itself. NN2 corresponds to NN in Fig.1 which illustrates our previous work. NN2 also plays the same role as NN. Training data for NN2 are obtained from point correspondences of a distinct marker point when the object undergoes a known rotation. During rotation, target object image irradiance $E_j(x, y)$ is acquired for each light source $j$. Suppose the target object is rotated along the horizontal axis (x-axis) and along the vertical axis (y-axis) with known step. In particular, suppose the target object is rotated in 5 degree steps from $-90$ degrees to $+90$ degrees along the horizontal axis and along the vertical axis. Images are acquired under each of four light sources. Fig.2 illustrates the overview of the proposed method.

As illustrated in Fig.2, a virtual sphere is generated through images during the rotation of the target object itself. The actual process to generate a virtual sphere image is to track the position of the marker point with high accuracy during two DOF rotation of the target object. However, the geometrical estimation of tracking of marker position is unpractical and inefficient for each rotated object pose. The problem is that it is difficult to track the marker position automatically and it is also a hard task to adjust the marker position for each rotated pose manually. In this approach, a neural network estimates the position of the marker as illustrated in Fig.3. This neural network corresponds to NN0 in Fig.2. This NN0 learns the mapping of the geometrically settled rotation angles of the marker to the $(x, y)$ coordinate (i.e., observed image coordinate) of the

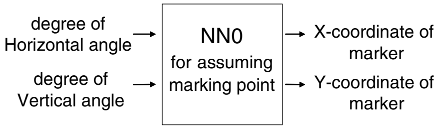


**Fig. 3.** neural network for assuming marker point

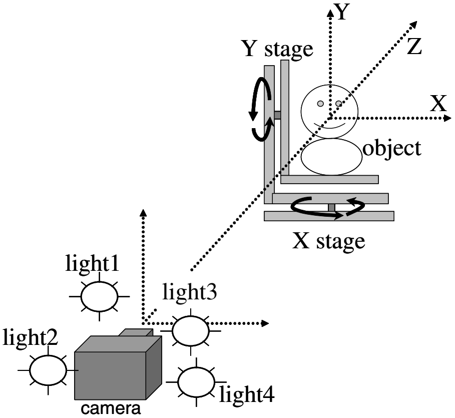


**Fig. 4.** Experimental environment

real position of the marker. Also, NN1 is used to obtain a smoothed images of the virtual sphere. NN1 learns and generalizes the mapping of the marker position $(x, y)$ to $(E_1, E_2, E_3, E_4)$.

Training data for NN0 are obtained from the correspondence between the object pose (degrees of horizontal angle and vertical angle) and the marker position $(x, y)$ through the object rotation with every 15 degree from $-90$ degrees to $+90$ degrees along the horizontal axis and along the vertical axis under each light source. Whereas, the target object is rotated with every 5 degrees in the horizontal and vertical rotation and $37 \times 37$ images are acquired through the rotation. NN0 is learned and generalized using a part of whole image data. Thus, it is not hard task to prepare the training data for NN0 manually and NN0 can be used to improve the accuracy of the marker position, that is, the virtual sphere image generation by NN1 can be also improved. This idea leads to have the confidential final NN2 for the classification of local surface.

## 4 Experimental Results

Fig.4 illustrates our experimental environment. The test object is placed on the rotation table as illustrated in Fig.2. Four light sources are used to illuminate the test object. One light source is aligned with the viewing direction. Four images are obtained under four different illuminations for each object pose during

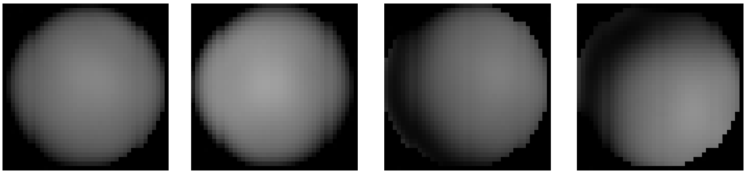


**Fig. 5.** Virtual sphere images under four light sources

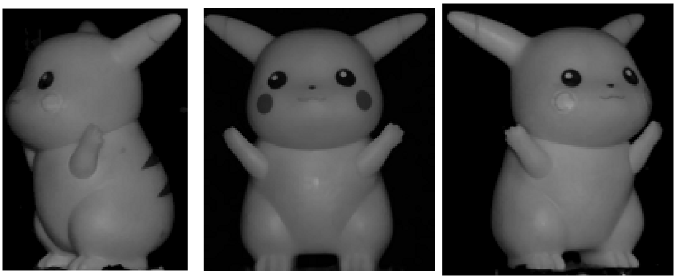


**Fig. 6.** Test object images with different poses

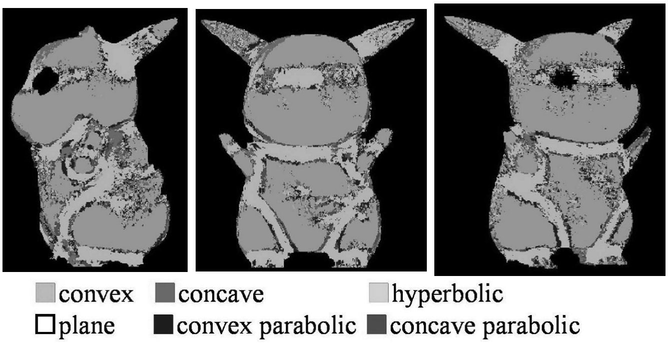


**Fig. 7.** Surface classification results

rotation. The test object is rotated with every 5 degrees between −90 degrees and +90 degrees. The total number of poses under rotation is 37 × 37, that is, a total of 37 × 37 × 4 images are taken to perform the self-calibration.

Each of Fig.5 shows the virtual sphere image of the test object illustrated in Fig.2. The virtual sphere images under four light sources were generated with 256 × 256 pixels by the neural networks (NN0 and NN1) in Fig.2. Images of a test object with different poses are shown in Fig.6. Fig.7 for each pose show the six local surface classes: convex, concave, convex parabolic, concave parabolic, hyperbolic surfaces and plane. Gray level is assigned according to the curvature class shown in the bottom of Fig.7. Curvature has similar values for the corresponding points at the different poses because it is view point independent. The results are qualitatively correct without using a distinct calibration object

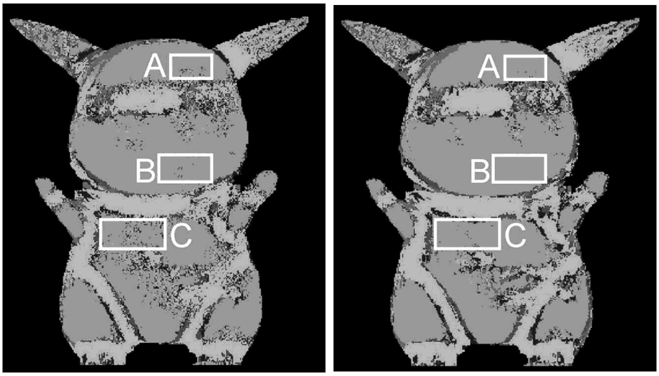


**Fig. 8.** Comparison of surface classifications

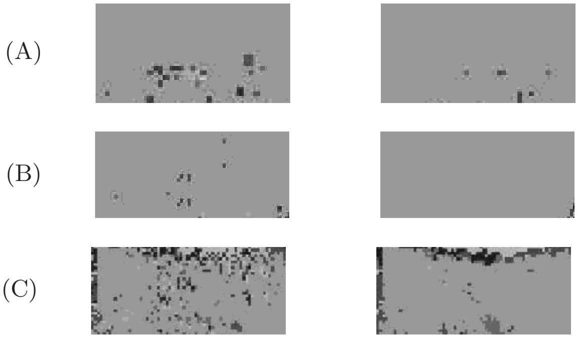


**Fig. 9.** Comparison of surface classification in the surrounding rectangulars in Fig.8

except in regions of cast shadow (around the ear regions) and black colored (eye regions).

Fig.8 shows the surface classification by the previous work [9] (left of Fig.8) and the proposed method (right of Fig.8).

In particular, Fig.9 shows the enlarged regions surrounded with rectangles A, B and C in Fig.8. The left column corresponds the left of Fig.8, and the right corresponds the right of Fig.8. It is observed that some errors or noises are reduced in comparison with the results of the previous work [9].

## 5 Conclusion

This paper demonstrates a method to classify the local surface curvature into six classes at each visible point on a test object without using any calibration object. Instead, two degrees of freedom rotation of the test object is used in self-calibration to generate a virtual sphere. When the virtual sphere images are generated, neural network for estimating the marker position is introduced.

This idea could increase the accuracy for the marker position, as a result, more robust results could be obtained in comparison with the previous approach. The proposed approach has the clear advantage for the improvement of generating a virtual sphere for neural network learning. This was observed through the experiments. Further subject are remained for the problem of cast shadow. Dealing with cast shadow and simplification of the implementation are remained as the future work.

## Acknowledgment

Iwahorifs research is supported by JSPS Grant-in-Aid for Scientific Research (C)(20500168), Chubu University Grant, Nitto Foundation. Wooodham's research is supported by the Natural Sciences and Engineering Research Council (NSERC).

## References

1. Woodham, R.J.: Gradient and curvature from the photometric stereo method, including local confidence estimation. Journal of the Optical Society of America, A 11, 3050–3068 (1994)
2. Iwahori, Y., Woodham, R.J., Bagheri, A.: Principal components analysis and neural network implementation of photometric stereo. In: Proc. IEEE Workshop on Physics-Based Modeling in Computer Vision, pp. 117–125 (1995)
3. Iwahori, Y., Woodham, R.J., Ozaki, M., Tanaka, H., Ishii, N.: Neural Network based Photometric Stereo with a Nearby Rotational Moving Light Source. IEICE Transactions on Information and Systems E80-D(9), 948–957 (1997)
4. Angelopoulou, E., Wolff, L.B.: Sign of Gaussian Curvature From Curve Orientation in Photometric Space. IEEE Trans. on PAMI 20(10), 1056–1066 (1998)
5. Okatani, T., Deguchi, K.: Determination of Sign of Gaussian Curvature of Surface from Photometric Data. Trans. of IPSJ 39(5), 1965–1972 (1998)
6. Iwahori, Y., Fukui, S., Woodham, R.J., Iwata, A.: Classification of Surface Curvature from Shading Images Using Neural Network. IEICE Trans. on Information and Systems E81-D(8), 889–900 (1998)
7. Iwahori, Y., Fukui, S., Fujitani, C., Woodham, R.J., Iwata, A.: Relative Magnitude of Gaussian Curvature from Shading Images Using Neural Network. In: Khosla, R., Howlett, R.J., Jain, L.C. (eds.) KES 2005. LNCS (LNAI), vol. 3681, pp. 813–819. Springer, Heidelberg (2005)
8. Chen, S., Cowan, C.F.N., Grant, P.M.: Orthogonal least squares learning algorithm for radial basis function networks. IEEE Transactions on Neural Networks 2(2), 302–309 (1991)
9. Ding, Y., Iwahori, Y., Nakamura, T., He, L., Woodham, R.J., Itoh, H.: Relative Magnitude of Gaussian Curvature Using Neural Network and Object Rotation of Two Degrees of Freedom. In: MVA 2007, IAPR Conference on Machine Vision Applications, pp. 110–113 (2007)

# Control of Hypothesis Space Using Meta-knowledge in Inductive Learning

Nobuhiro Inuzuka, Hiroyuki Ishida, and Tomofumi Nakano

Nagoya Institute of Technology,
Gokiso-cho Showa, Nagoya 466-8555, Japan
{inuzuka,nakano.tomfumi}@nitech.ac.jp, raituti@phaser.elcom

**Abstract.** Inductive logic programming (ILP) is effective for classification learning because it constructs hypotheses combining background knowledge. On the other hand it makes the cost of search for hypothesis large. This paper proposes a method to prune hypothesis using a kind of semantic knowledge. When an ILP system uses a top-down search, after it visits a clause (rule) it explore another clause by adding a condition. The added condition may be redundant with other conditions in the clause or the condition may causes the body of clause unsatisfied. We study to represent and use to treat the redundancy and unsatisfactory of conditions as meta-knowledge of predicates. In this paper we give a formalism of meta-knowledge and show to use it with an ILP algorithm. We also study a method to generate meta-knowledge automatically. The method generates meta-knowledge which controls redundancy and contradiction with respect to predicates by testing properties extensionally.

## 1 Introduction

*Inductive learning* (IL) is to make hypothetic theory to explain phenomena or data. This is important to help scientific discovery by computers and to support decision making or engineering. Learning is also a large topic of AI. IL has been an important area in academic and also practical sense. Since IL involved the framework of logic programming (LP) in 90's, it gained a feature as a combined area of logic and computation [1]. It is called *inductive logic programming (ILP).* While LP uses deduction for programming ILP uses induction.

A typical ILP problem is classification. It gives a hypothesis as a logic program to explain data, called examples. When we have a set of examples (or *positive examples*) to be explained, anti-examples (or *negative examples*) which must not be explained, and some relevant knowledge, *background knowledge*, to our domain, the task is to give a set of rules (a *hypothesis*) which explains all examples and excludes anti-examples by combining the background knowledge. Combining the background knowledge means to construct rules by using conditions given as predicates in background knowledge.

A useful ILP classification algorithm is greedy and top-down. A representative is the classical algorithm FOIL[2,3]. It continues to find a clause to cover some examples and no anti-examples until all examples are covered by at least a clause.

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 911–918, 2008.

A clause is constructed by adding literals to an initial empty clause. Literals play a role as conditions to exclude anti-examples.

The algorithm is useful but expensive when background knowledge allows many choices of conditions to be added. Our contribution is to prune possible conditions using semantic knowledge. ILP algorithms have been given functions of controlling hypotheses to overcome this combinatorial problem. Language bias to allow syntax of rules is well studied and used in many ILP systems including Progol[4,5] and Warmr[6]. The system given in [7] shares our aim by learning knowledge about functional dependency and symmetiricity on predicates in background knowledge. Our formalism treats related predicates as well and large range of knowledge based on redundancy and contradiction.

## 2 Inductive Logic Programming

A rule is a clause, a formula of the form $\forall(h \leftarrow b_1 \wedge \ldots \wedge b_n)$ for literals $h, b_1, \ldots, b_n$. When we have a set of examples $\mathcal{E}^+$ to be explained, anti-examples $\mathcal{E}^-$ which must not be explained, and some relevant knowledge, *background knowledge*, denoted by $\mathcal{B}$, the task is to give a set of clauses (a *hypothesis*) $H$ s.t. $\mathcal{B} \cup H \vdash e^+$ for all $e^+ \in \mathcal{E}^+$ and $B \cup H \not\vdash e^-$ for all $e^- \in \mathcal{E}^-$. $\vdash$ means the logical derivation. When $\mathcal{B} \cup H \vdash e^+$, $H$ is said to cover $e$. The task is to find $H$ which covers all examples and no anti-examples.

Among many classification methods in ILP framework the top-down algorithm shown in Fig. 1 is representative. It constructs clauses in a greedy way using an appropriateness, such as an information gain criteria for FOIL[2,3].

The time complexity of the algorithm depends on two factors; one is the number of examples and the other is the number of possible hypotheses. Each time to test if a hypothesis covers examples takes a cost linear to the number of examples. Although the number of possible hypotheses affects the cost also in learner, it grows rapidly by the number of predicates. It is worth studying to prune hypotheses without covering test for scalability of ILP algorithms.

## 3 Meta-knowledge and Pruning Using Meta-knowledge

**Redundant Hypotheses and Contradictory Hypotheses.** Think of predicates $\mathsf{inc}(A, B)$ meaning $A + 1 = B$ and $\mathsf{dec}(A, B)$ meaning $A - 1 = B$ for integers, and try to make a hypothesis using them. If a clause includes $\mathsf{inc}(A, B)$ as a condition, adding $\mathsf{dec}(B, A)$ to the clause is no meaning because $\mathsf{inc}(A, B)$ implies $\mathsf{dec}(B, A)$. That is, $\mathsf{inc}(A, B) \wedge \mathsf{dec}(B, A)$ is redundant. On the other hand $\mathsf{inc}(A, B) \wedge \mathsf{inc}(B, A)$ makes contradiction because of the asymmetricity of $\mathsf{inc}$. When $\mathsf{inc}(A, B)$ is true, $A$ is smaller than $B$, denoted by $\mathsf{order}(A, B)$. It is an order relation. $\mathsf{dec}(A, B)$ is also related with the order and implies $\mathsf{order}(B, A)$. Because $\mathsf{order}$ is transitive and asymmetric, a clause that includes $\mathsf{inc}(A, B) \wedge \mathsf{inc}(B, C)$ can not have $\mathsf{dec}(A, C)$.

```
Input   $\mathcal{E}^+$ : a set of positive examples of the target relation $r$,
        $\mathcal{E}^-$ : a set of negative examples of the target relation $r$,
        $\mathcal{B}$  : background knowledge (a theory);
Output  $H$  : a theory (hypothesis) which covers all examples in $\mathcal{E}^+$ and
             no examples in $\mathcal{E}^-$wrt $\mathcal{B}$.
$H := \emptyset$;
while $\mathcal{E}^+ \neq \emptyset$ do
        $c :=$ "$r(A, B, \ldots) \leftarrow .$";
        while $\{e \in \mathcal{E}^- \mid \mathcal{B} \cup H \cup \{c\} \models e\} \neq \emptyset$ do
            $\mathcal{L} :=$ the set of all candidate literals to add $c$;
            Add the most appropriate literal from $L$ to the clause $c$;
        $H := H \cup \{c\}$;
$\mathcal{E}^+ := \mathcal{E}^+ - \{e \in \mathcal{E}^{+\prime} | \mathcal{B} \cup H \cup \{c\} \models e\}$.
```

**Fig. 1.** The top-down clarification algorithm

We understand the redundancy and contradiction from semantic properties, such as the asymmetricity of $\mathsf{inc}$. Such properties are not only of predicates in $\mathcal{B}$, but of predicates to describe potential properties, such as transitivity of $\mathsf{order}$.

**Meta-Knowledge and its Formalization.** As we discussed in the previous paragraph $\mathsf{inc}$ has properties and we can use them to eliminate hypotheses. The redundancy of $\mathsf{inc}(A, B) \land \mathsf{dec}(B, A)$ is because of functional dependency from $\mathsf{inc}(A, B)$ to $\mathsf{dec}(B, A)$. (In this case we have the functional dependency also for the converse direction.) The functional dependency is represented by, $\mathsf{inc}(A, B) \rightarrow \mathsf{dec}(B, A)$. On the other hand, the asymmetricity of $\mathsf{inc}(A, B)$ is represented by, $\mathsf{inc}(A, B) \land \mathsf{inc}(B, A) \rightarrow \square$, where $\square$ denotes the contradiction. We may understand these formulae meta-knowledge.

When we use $\mathsf{inc}(A, B)$ and $\mathsf{dec}(C, D)$ values for variables are in range of numbers and $\mathsf{inc}(A, B)$ (and $\mathsf{dec}(C, D)$) keeps the order $A < B$ (and $C > D$, resp.) on the range. That is, for $\mathsf{inc}(A, B)$ there is an embedded relation $\mathsf{emb}_{num}(A, B)$[1] and it satisfies the properties of orders. That is, it can be represented by,

$$\mathsf{inc}(A, B) \rightarrow \mathsf{emb}_{num}(A, B), \qquad \mathsf{dec}(A, B) \rightarrow \mathsf{emb}_{num}(B, A),$$
$$\mathsf{emb}_{num}(A, B) \land \mathsf{emb}_{num}(B, C) \rightarrow \mathsf{emb}_{num}(A, C),$$
$$\mathsf{emb}_{num}(A, B) \land \mathsf{emb}_{num}(B, A) \rightarrow \square.$$

The first two connect $\mathsf{inc}$ and $\mathsf{dec}$ to $\mathsf{emb}_{num}$ and the third and forth formulae represent the transitive and asymmetric properties of $\mathsf{emb}_{num}$. To represent meta-knowledge we need related predicates like $\mathsf{emb}_{num}$.

For a background knowledge $\mathcal{B}$, predicate symbols in $\mathcal{B}$ and potential predicate symbols related to $\mathcal{B}$ are called *related predicate symbols* to $\mathcal{B}$. For example when $\mathcal{B}$ includes $\mathsf{inc}$ and $\mathsf{dec}$ the related predicate symbols include $\mathsf{emb}_{num}$ as well. The intended interpretation $I_{\mathcal{B}}$ of the related predicates is assumed. We define

[1] It correspond to the relation $\mathsf{order}$ in the previous paragraph.

**Table 1.** Meta-knowledges and their representation

| meta-knowledge | explanation | representation |
|---|---|---|
| types | variable $X$ in $p(X, \cdots)$ is of the type $A$ | $p(X, \cdots) \rightarrow \mathsf{typeA}(X)$<br>$\mathsf{typeA}(X) \wedge \mathsf{typeB}(X) \rightarrow \Box$ |
| functional dependency | if $p(X, \ldots)$ then $q(X, \ldots)$ | $p(X, \ldots) \rightarrow q(X, \ldots)$ |
| symmetricity | $p(X, Y)$ is symmetric | $p(X, Y) \rightarrow p(Y, X)$ |
| asymmetricity | $p(X, Y)$ is asymmetric | $p(X, Y) \wedge p(Y, X) \rightarrow \Box$ |
| transitivity | $\langle X, Y \rangle$ is transitive in $p(X, Y, \cdots)$ | $p(X, Y, \cdots) \wedge p(Y, Z, \cdots)$ $\rightarrow p(X, Z, \cdots)$ |
| reflexivity | $p(X, Y)$ is reflexive | $p(X, X)$ |
| irreflexivity | $p(X, Y)$ is irreflexive | $p(X, X) \rightarrow \Box$ |
| uniqueness | $Y$ is unique for $X$ in $p(X, Y)$ | $p(X, Y) \wedge p(X, Z)$ $\wedge Y \neq Z \rightarrow \Box$ |
| embedded transitivity | a transitive relation $\mathsf{emb}_p(X, Y)$ is embedded in $p(X, Y, \cdots)$ | $p(X, Y, \cdots) \rightarrow \mathsf{emb}_p(X, Y)$<br>$\mathsf{emb}_p(X, Y) \wedge \mathsf{emb}_p(Y, Z)$ $\rightarrow \mathsf{emb}_p(X, Z)$ |
| other properties of embedded relations | similarly represented as a formula embedding a predicate $p$ to a related predicates $\mathsf{emb}_p$ and other formulae to describe the property of $\mathsf{emb}_p$. | |

a formula $\mu$ as a *meta-knowledge* of $\mathcal{B}$ when $\mu$ includes no constant symbols and no function symbols and satisfies $I_{\mathcal{B}} \models \mu$, where $\models$ means satisfaction.

Table 1 shows major meta-knowledge. It includes major properties of binary relations. It also includes uniqueness of an argument for another.

When a predicate is embedded in other relation which has some property it is represented by at least two formulae. The first embeds the predicated to the embedded relation and the others represent the property. Constraint by the type of arguments can also be represented as meta-knowledge using a related predicate $\mathsf{typeA}$ representing that a value is in a type TypeA. The meta-knowledge for type constraints consists of a formula connecting a predicate with an argument of TypeA to $\mathsf{typeA}$ and formulae representing exclusion or inclusion among types.

**Elimination Strategy Using Meta-Knowledge.** In order to describe a theorem which is a base of our control strategy of hypothesis space we introduce a denotation. A bold italic letter, such as $\boldsymbol{x}$, denotes a tuple of variables. A formula $f$ (a literal $l$) including only variables in $\boldsymbol{x}$ is denoted by $f(\boldsymbol{x})$ ($l(\boldsymbol{x})$, resp). Note that $f(\boldsymbol{x})$ and $l(\boldsymbol{x})$ do not necessarily include all variables in $\boldsymbol{x}$. $f(\boldsymbol{c_x})$ ($l(\boldsymbol{c_x})$, resp) means a formula replaced each variable by a distinct new constant.

Then we have a theorem as follows:

**Theorem 1.** *Let $\mathcal{M}$ be a set of mete-knowledge of $\mathcal{B}$ and $l_1(\boldsymbol{x}), \cdots, l_n(\boldsymbol{x}), m(\boldsymbol{x})$ literals without constant and function symbols. Then if it holds*

$$\mathcal{M} \cup \{l_1(\boldsymbol{c_x}), \cdots, l_n(\boldsymbol{c_x})\} \vdash m(\boldsymbol{c_x}), \tag{1}$$

*the formula $l_1(\boldsymbol{x}) \wedge \cdots \wedge l_n(\boldsymbol{x}) \rightarrow m(\boldsymbol{x})$ is also a meta-knowledge of $\mathcal{B}$.*

*Proof.* When we apply the deduction theorem to formula (1) we have, $\mathcal{M} \vdash l_1(\boldsymbol{c_x}) \wedge \cdots \wedge l_n(\boldsymbol{c_x}) \rightarrow m(\boldsymbol{c_x})$. The constants in $\boldsymbol{c_x}$ can be seen as free variables because $\mathcal{M}$ includes no constants and no function symbols. Then by introducing a universal quantifier we have, $\mathcal{M} \vdash \forall \boldsymbol{x}(l_1(\boldsymbol{x}) \wedge \cdots \wedge l_n(\boldsymbol{x}) \rightarrow m(\boldsymbol{x}))$. Since $\mathcal{M}$ is meta-knowledge, i.e. $I_{\mathcal{B}} \models \mathcal{M}$ then it holds $I_{\mathcal{B}} \models \forall \boldsymbol{x}(l_1(\boldsymbol{x}) \wedge \cdots \wedge l_n(\boldsymbol{x}) \rightarrow m(\boldsymbol{x}))$, which yields the theorem. □

We have the following corollary directly from the theorem.

**Corollary 1 (elimination by contradiction or redundancy).** *Let $\mathcal{M}$ is a set of meta-knowledge and $f(\boldsymbol{x}) = l_1(\boldsymbol{x}) \wedge \ldots \wedge l_n(\boldsymbol{x})$ a conjunctive formula, $m(\boldsymbol{x})$ a literal. Then,*

1. *if it holds $\mathcal{M} \cup \{l_1(\boldsymbol{c_x}), \ldots l_n(\boldsymbol{c_x})\} \vdash m(\boldsymbol{c_x})$ then*

$$\{\boldsymbol{x} \in |\mathcal{I}|^n \mid \mathcal{I} \models f(\boldsymbol{x})\} = \{\boldsymbol{x} \in |\mathcal{I}|^n \mid \mathcal{I} \models (f \wedge l)(\boldsymbol{x})\};$$

2. *if it holds $\mathcal{M} \cup \{l_1(\boldsymbol{c_x}), \ldots l_n(\boldsymbol{c_x}), m(\boldsymbol{c_x})\} \vdash \square$ then*

$$\{\boldsymbol{x} \in |\mathcal{I}|^n \mid \mathcal{I} \models (f \wedge l)(\boldsymbol{x})\} = \emptyset.$$

As a result we have a control strategy. Given $\mathcal{M}$ on an exploration of a clause $c = t(\boldsymbol{x}) \leftarrow l_1(\boldsymbol{x_1}) \wedge \cdots \wedge l_n(\boldsymbol{x_n})$, if it holds at least one of (1) $\mathcal{M} \cup \{l_1(\boldsymbol{c_x}), \cdots, l_n(\boldsymbol{c_x})\} \vdash m(\boldsymbol{c_x})$ and (2) $\mathcal{M} \cup \{l_1(\boldsymbol{c_x}), \cdots, l_n(\boldsymbol{c_x}), m(\boldsymbol{c_x})\} \vdash \square$, the clause $c' = t(\boldsymbol{x}) \leftarrow l_1(\boldsymbol{x}) \wedge \cdots \wedge l_n(\boldsymbol{x}) \wedge m(\boldsymbol{x})$ need not be explored. The strategy sees only the derivation among ground (no variable) literals in the clause by replacing variables of literals by constants.

**An Example to Use Meta-Knowledge.** Let us think of a clause $c = \mathsf{add}(A, B) \leftarrow \mathsf{dec}(A, C)$ and a literal $\mathsf{inc}(A, C)$. As a related meta-knowledge we take

$$\begin{aligned}\mathcal{M}_1 = \{\ & \mathsf{dec}(x_1, x_2) \rightarrow \mathsf{emb}_{num}(x_1, x_2),\ \ \mathsf{dec}(x_1, x_2) \rightarrow \mathsf{emb}_{num}(x_2, x_1),\\ & \mathsf{emb}_{num}(x_1, x_2) \wedge \mathsf{emb}_{num}(x_2, x_1) \rightarrow \square\ \}.\end{aligned}$$

To use Corollary 1 $\mathsf{dec}(A, C)$ and $\mathsf{inc}(A, C)$ are transformed to ground $\mathsf{inc}(\boldsymbol{c_A}, \boldsymbol{c_C})$ and $\mathsf{dec}(\boldsymbol{c_A}, \boldsymbol{c_C})$. Then we can easily see,

$$\mathcal{M}_1 \cup \{\mathsf{inc}(\boldsymbol{c_A}, \boldsymbol{c_C}), \mathsf{dec}(\boldsymbol{c_A}, \boldsymbol{c_C})\} \vdash \square$$

The clause $\mathsf{add}(A, B) \leftarrow \mathsf{dec}(A, C) \wedge \mathsf{inc}(A, C)$ is found not to be explored.

**Table 2.** Meta-knowledge on predicates $\mathsf{memb}$ and $\mathsf{cmp}$.

| | | |
|---|---|---|
| $\mathsf{memb}(A,\_) \rightarrow \mathsf{elem}(A)$, | $\mathsf{memb}(\_,A) \rightarrow \mathsf{list}(A)$, | $\mathsf{cmp}(A,\_,\_) \rightarrow \mathsf{list}(A)$, |
| $\mathsf{cmp}(\_,A,\_) \rightarrow \mathsf{elem}(A)$, | $\mathsf{cmp}(\_,\_,A) \rightarrow \mathsf{list}(A)$, | $\mathsf{elem}(A) \wedge \mathsf{list}(A) \rightarrow \square$, |
| $\mathsf{cmp}(A,\_,B) \rightarrow \mathsf{emb}_{\mathsf{list}}(A,B)$, | | $\mathsf{emb}_{\mathsf{list}}(A,B) \wedge \mathsf{emb}_{\mathsf{list}}(B,A) \rightarrow \square$, |
| $\mathsf{emb}_{\mathsf{list}}(A,B) \wedge \mathsf{emb}_{\mathsf{list}}(B,C) \rightarrow \mathsf{emb}_{\mathsf{list}}(A,C)$, | | |
| $\mathsf{cmp}(A,B,\_) \wedge \mathsf{cmp}(A,C,\_) \wedge B \neq C \rightarrow \square$, | $\mathsf{cmp}(A,\_,B) \wedge \mathsf{cmp}(A,\_,C) \wedge B \neq C \rightarrow \square$ | |

We implemented the ILP algorithm in Fig.1 with the control strategy. Table 2 shows meta-knowledge for an induction of $\mathsf{memb}(x,y)$, testing if $x$ is a member of a list $y$, using $\mathsf{cmp}(x,y,z)$, which decomposes a list $x$ to its head $y$ and the tail $z$. It includes related predicates $\mathsf{elem}$, $\mathsf{list}$, $\mathsf{emb}_{\mathsf{list}}$ as well as $\mathsf{memb}$ and $\mathsf{cmp}$. The induction process successfully worked using the meta-knowledge. While the induction without the strategy tested 29 clauses for the first three conditions of a rule, the method explored only 10 clauses.

## 4 Induction of Meta-knowledge for Induction

**Acquiring Meta-Knowledge.** As we discussed there are wide range of meta-knowledge because they are characterised by just true formulae. It is possible there are many unaware other meta-knowledge than we presented. There are three approaches to prepare meta-knowledge for a domain.

*To declare names of meta-knowledge in background knowledge description.* For example, uniqueness, asymmetricity and embedded transitivity should be declared for $\mathsf{inc}$ as well as for its argument types. This approach is practical and easily combined with ILP systems that use language bias declaration. There are two disadvantages; it charges the responsibility of declaration to users and the complication of ILP background knowledge description goes worse.

*To reason meta-knowledge from known meta-knowledge.* When a predicate is defined by a conjunction of other several predicates among which at least one of the predicates is irreflexive, the defined predicate is also irreflexive. Using this kind of rules we may construct a system reasoning meta-knowledge. Rules are given by rigor set-theoretic algebra. However, they do not cover all of meta-knowledge. To establish reasoning system is interesting but far from the practice.

*To induce meta-knowledge from extensional database.* We have extensional knowledge-base. Even if we do not have, extensional tuples for a specific domain can be calculated. To induce meta-knowledge from extensional DB is an idea. It is a defect that extensional DBs are required also for implicit related predicates. However usually related predicates are natural like the order relation or types for domains. Automatic preparation of meta-knowledge brings large merit to users. A possibility to find unaware knowledge is also another merit. This is challenging and a reasonable medium between the previous two approaches.

**Meta-Knowledge Induction Algorithm.** Induction from extensional DB is an interesting and practical solution. The proposing algorithm (Fig.2) explores

```
Input:  Pred : a set of related predicate symbols to a background knowledge DB,
        $D$    : a closed finite domain for predicates,
        $\mathcal{I}$    : extensional database or intended interpretation of Pred,
               the database which is given within the domain $D$,
Output  Meta : A set of meta-knowledge.
1. OPEN := $\emptyset$
2. for each $p/n \in$ Pred do add $p(X_1, \ldots, X_n)$ to OPEN;
3. while OPEN $\neq \emptyset$ do
4.     pick an element $f(\boldsymbol{x}) \in$ OPEN;
5.     for each a syntactically legal literal $l$ to add $f$ do
6.         if $\{\boldsymbol{x} \in D^{|\boldsymbol{x}|} \mid \mathcal{I} \models f(\boldsymbol{x})\} = \{\boldsymbol{x} \in D^{|\boldsymbol{x}|} \mid f(\boldsymbol{x}) \wedge l(\boldsymbol{x})\}$ then add $f \to l$ to $\mathcal{M}$;
7.         if $\{\boldsymbol{x} \in D^{|\boldsymbol{x}|} \mid \mathcal{I} \models f(\boldsymbol{x}) \wedge l(\boldsymbol{x})\} = \emptyset$ then add $f \wedge l \to \square$ to $\mathcal{M}$;
8.         else add $f \wedge l$ to OPEN.
```

**Fig. 2.** Induction algorithm of meta-knowledge

exhaustively meta-knowledge in clausal form. It takes a set of predicates symbols Pred including all related predicates and constructs formulae as meta-knowledge. It also takes a closed finite domain $D$ for predicates. Normally the domain is infinite but if we test formulae extensionally the domain need be finite. We restrict some closed domain. A domain $D$ is closed if we apply some values in $D$ for a predicate the predicate does not give values outside of $D$. A plain induction algorithm FOIL also requires extensional DB in a closed domain. An extensional DB for predicates is also given inside of the closed domain.

In the algorithm OPEN keeps a conjunction $f$ of literals and algorithm explores a literal $l$ which is extensionally entailed from $f$ (line 6), or a literal $l$ with which the conjunction, i.e. $f \wedge l$, makes contradiction extensionally (line 7). If neither case is satisfied, the new conjunction $f \wedge l$ is registered in OPEN for further exploration. As in a normal-ILP induction algorithm language bias including the type constraint can be used to prune many meta-knowledge candidates.

**An Experiment and its Result.** The algorithm was implemented and examined. In an experiment we used predicates $\mathsf{add}$ and $\mathsf{sub}$ for addition and subtraction, $\mathsf{emb}_{num}$ for the embedded order on numbers, $\mathsf{eq}_{num}$ and $\mathsf{neq}_{num}$ for equality and inequality of numbers.

Many meta-knowledge formulae are produced. The number of the formulae produced was counted for each of size of the conjunction $f$. The run-time was also measured. For $f$ of length one it produced 208 formulae for meta-knowledges and took 0.21 seconds. For $f$ of length two it produced 23860 formulae and took 38.66 seconds. Table 3 shows a part of formulae produced.

The result validates the algorithm. It produces unexpectedly many formulae but the execution time was reasonable. The algorithm has a problem that produced formulae includes redundancy. It means that some knowledge can be implied from others. To have compact knowledge set we may consider a pruning technique for meta-knowledge induction as we developed for plain induction.

**Table 3.** A part of meta-knowledge induced by the algorithm

| | |
|---|---|
| $\mathsf{add}(A, \_, B) \rightarrow \mathsf{emb}_{num}(A, B)$ | $\mathsf{add}(\_, A, B) \rightarrow \mathsf{emb}_{num}(A, B)$ |
| $\mathsf{sub}(A, B, \_) \rightarrow \mathsf{emb}_{num}(B, A)$ | $\mathsf{sub}(A, \_, B) \rightarrow \mathsf{emb}_{num}(B, A)$ |
| $\mathsf{add}(A, B, C) \rightarrow \mathsf{sub}(C, A, B)$ | $\mathsf{add}(A, B, C) \rightarrow \mathsf{sub}(C, B, A)$ |
| $\mathsf{sub}(A, B, C) \rightarrow \mathsf{sub}(A, C, B)$ | $\mathsf{add}(A, B, C) \rightarrow \mathsf{add}(B, A, C)$ |
| $\mathsf{sub}(A, B, C) \rightarrow \mathsf{add}(B, C, A)$ | $\mathsf{sub}(A, B, C) \rightarrow \mathsf{add}(C, B, A)$ |
| $\mathsf{eq}_{num}(A, \_) \rightarrow \mathsf{eq}_{num}(A, A)$ | $\mathsf{eq}_{num}(A, B) \rightarrow \mathsf{eq}_{num}(B, A)$ |
| $\mathsf{sub}(A, \_, B) \rightarrow \mathsf{neq}_{num}(B, A)$ | $\mathsf{emb}_{num}(A, B) \rightarrow \mathsf{neq}_{num}(A, B)$ |
| $\mathsf{emb}_{num}(A, B) \rightarrow \mathsf{neq}_{num}(B, A)$ | $\mathsf{neq}_{num}(A, B) \rightarrow \mathsf{neq}_{num}(B, A)$ |
| $\mathsf{add}(A, \_, B) \wedge \mathsf{add}(A, B, A) \rightarrow \square$ | $\mathsf{add}(A, C, B) \wedge \mathsf{add}(A, B, C) \rightarrow \square$ |
| $\mathsf{add}(A, B, C) \wedge \mathsf{sub}(A, B, C) \rightarrow \square$ | $\mathsf{add}(A, \_, \_) \wedge \mathsf{emb}_{num}(A, A) \rightarrow \square$ |
| $\mathsf{add}(B, \_, A) \wedge \mathsf{emb}_{num}(A, B) \rightarrow \square$ | $\mathsf{add}(\_, B, A) \wedge \mathsf{emb}_{num}(A, B) \rightarrow \square$ |
| $\mathsf{emb}_{num}(B, A) \wedge \mathsf{emb}_{num}(A, B) \rightarrow \square$ | $\mathsf{eq}_{num}(A, B) \wedge \mathsf{neq}_{num}(A, B) \rightarrow \square$ |
| $\mathsf{neq}_{num}(A, \_) \wedge \mathsf{neq}_{num}(A, A) \rightarrow \square$ | |

## 5 Conclusion

We studied a technique of controlling hypotheses space for ILP classification algorithms. We formalized meta-knowledge and gave the theorem to use them based on redundancy and contradiction. The pruning using meta-knowledge is very effective and it treats wide range of constraints. Not only to the top-down induction algorithm, the technique can be applied to most of algorithms to treat conjunctive conditions in hypotheses. The extensional induction of meta-knowledge was also investigated. Although the technique needs further study, the experiments showed that the approach is promising.

## References

1. Muggleton, S.: Inductive Logic Programming. Academic Press, London (1992)
2. Quinlan, J.R.: Learning Logical Definitions from Relations. Machine Learning 5, 239–266 (1990)
3. Quinlan, J.R., Cameron-Jones, R.M.: FOIL: A Midterm Report. In: Brazdil, P.B. (ed.) ECML 1993. LNCS, vol. 667, pp. 3–20. Springer, Heidelberg (1993)
4. Muggleton, S.: Inverting Entailment and Progol. Machine Intelligence 14, 135–190 (1993)
5. Muggleton, S.: Inverse Entailment and Progol. New Generation Computing 13(3-4), 245–286 (1995)
6. Dehaspe, L., De Raedt, L.: Mining Association Rules in Multiple Relations. In: Džeroski, S., Lavrač, N. (eds.) ILP 1997. LNCS, vol. 1297, pp. 125–132. Springer, Heidelberg (1997)
7. McCreath, E., Sharma, A.: Extraction of Meta-Knowledge to Restrict the Hypothesis Space for ILP Systems. In: 8th Australian Joint Conf. on AI, pp. 75–82 (1995)

# On Capacity of Memory in Chaotic Neural Networks with Incremental Learning

Toshinori Deguchi[1], Keisuke Matsuno[1], and Naohiro Ishii[2]

[1] Gifu National College of Technology
[2] Aichi Institute of Technology

**Abstract.** Neural networks are able to learn more patterns with the incremental learning than with the correlative learning. The incremental learning is a method to compose an associate memory using a chaotic neural network. In former work, it was found that the capacity of the network increases along with its size, which is the number of the neurons in the network, until some threshold size and that it decreases over that size. The threshold size and the capacity varied between 2 different learning parameters. In this paper, the capacity of the networks was investigated changing the learning parameter. Through the computer simulations, it turned out that the capacity also increases in proportion to the network size in larger sizes and that the capacity of the network with the incremental learning is above 11 times larger than the one with correlative learning.

## 1 Introduction

The incremental learning proposed by the authors is highly superior to the auto-correlative learning in the ability of pattern memorization[1,2].

The idea of the incremental learning is from the automatic learning[3]. In the incremental learning, the network keeps receiving the external inputs. If the network has already known a input pattern, it recalls the pattern. Otherwise, each neuron in it learns the pattern gradually.

The neurons used in this learning are the chaotic neurons, and their network is the chaotic neural network, which was developed by Aihara[4].

In former work, we investigated the capacity of the networks[5]. Through the simulations, we found that the capacity of the network grows up along with its size, which is the number of neurons in the network, until some threshold size and that it falls off over the size. The threshold size and the capacity varied between 2 different learning parameters.

In this paper, first, we explain the chaotic neural networks and the incremental learning and refer to the former work on the capacities at the 2 learning parameters[5], then examine the maximum capacity of the network with simulations changing the learning parameter and show that the capacity is also in proportion to the network size in larger sizes with appropriate parameter.

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 919–925, 2008.

## 2 Chaotic Neural Networks and Incremental Learning

The incremental learning was developed by using the chaotic neurons. The chaotic neurons and the chaotic neural networks were proposed by Aihara[4].

The incremental learning provides an associative memory and the network is an interconnected network, in which each neuron receives one external input, and is defined as follows:

$$x_i(t+1) = f\big(\xi_i(t+1) + \eta_i(t+1) + \zeta_i(t+1)\big) \ , \tag{1}$$

$$\xi_i(t+1) = k_s\xi_i(t) + \upsilon A_i(t) \ , \tag{2}$$

$$\eta_i(t+1) = k_m\eta_i(t) + \sum_{j=1}^{n} w_{ij}x_j(t) \ , \tag{3}$$

$$\zeta_i(t+1) = k_r\zeta_i(t) - \alpha x_i(t) - \theta_i(1-k_r) \ , \tag{4}$$

where $x_i(t+1)$ is the output of the $i$-th neuron at time $t+1$, $f$ is the output sigmoid function described below in (5), $k_s, k_m, k_r$ are the time decay constants, $A_i(t)$ is the input to the $i$-th neuron at time $t$, $\upsilon$ is the weight for external inputs, $n$ is the size—the number of the neurons in the network, $w_{ij}$ is the connection weight from the neuron $j$ to the neuron $i$, and $\alpha$ is the parameter that specifies the relation between the neuron output and the refractoriness.

$$f(x) = \frac{2}{1+\exp(\frac{-x}{\varepsilon})} - 1 \ . \tag{5}$$

The parameters in the chaotic neurons are assigned to the values in Table 1 in our works[1,2,5].

In the incremental learning, the network has each pattern inputted during fixed steps—it is 50 steps in this paper—before moving to the next one. After all the patterns are inputted, the first pattern comes repeatedly. In this paper, a set is defined as a period through all patterns inputted from the first pattern to the last pattern.

During the learning, a neuron which satisfies the condition of (6) changes the connection weights as in (7)[1].

$$\xi_i(t) \times (\eta_i(t) + \zeta_i(t)) < 0 \ . \tag{6}$$

**Table 1.** Parameters

| |
|---|
| $\upsilon = 2.0,$ |
| $k_s = 0.95,$ |
| $k_m = 0.1,$ |
| $k_r = 0.95,$ |
| $\alpha = 2.0,$ |
| $\theta_i = 0,$ |
| $\varepsilon = 0.015$ |

$$w_{ij} = \begin{cases} w_{ij} + \Delta w, & \xi_i(t) \times x_j(t) > 0 \\ w_{ij} - \Delta w, & \xi_i(t) \times x_j(t) \leq 0 \end{cases} \quad (i \neq j) \ , \tag{7}$$

where $\Delta w$ is the learning parameter.

In this learning, the initial values of the connection weights can be 0, because some of the neurons' outputs are changed by their external inputs and this makes the condition establish in some neurons. Therefore, all initial values of the connection weights are set to be 0 in this paper. $\xi_i(0)$, $\eta_i(0)$, and $\zeta_i(0)$ are also set to be 0.

To confirm that the network has learned a pattern after the learning, the pattern is inputted to the usual Hopfield's type network which have the same connection weights as the chaotic neural network. That the Hopfield's type network with the connection weights has the pattern in its memory has the same meaning that the chaotic neural network recalls the pattern quickly when the pattern inputted. Therefore, it is the convenient way to use the Hopfield's type network to check the success of the learning.

## 3 Capacity

In this section, we retrace the simulations in the former work[5]. In the simulations, we settled the learning parameter $\Delta w$ to be 0.05 which was used in the former works[1,2]. The simulations investigated the number of success, which means the number of patterns that the network learned in it successfully, after 50 sets of learning in the networks composed of 50, 100, 200, 300, or 400 neurons.

In each network, the number of patterns to be learned moved from 10 to 300. These patterns are the random patterns generated with the method that all

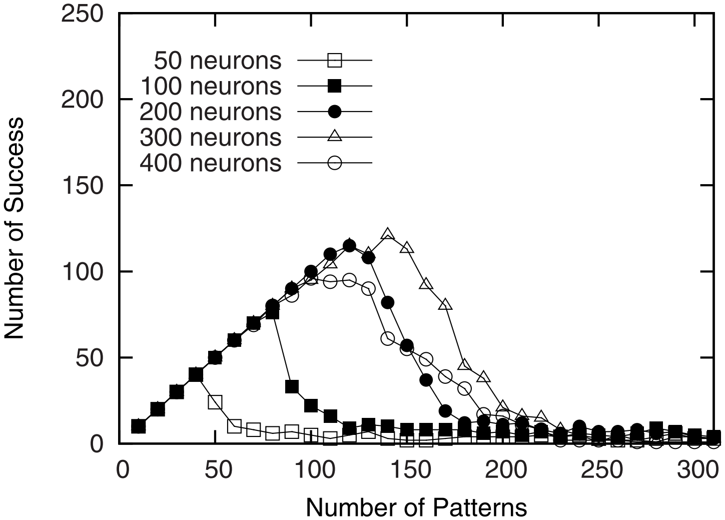


**Fig. 1.** Number of learned patterns ($\Delta w = 0.05$)

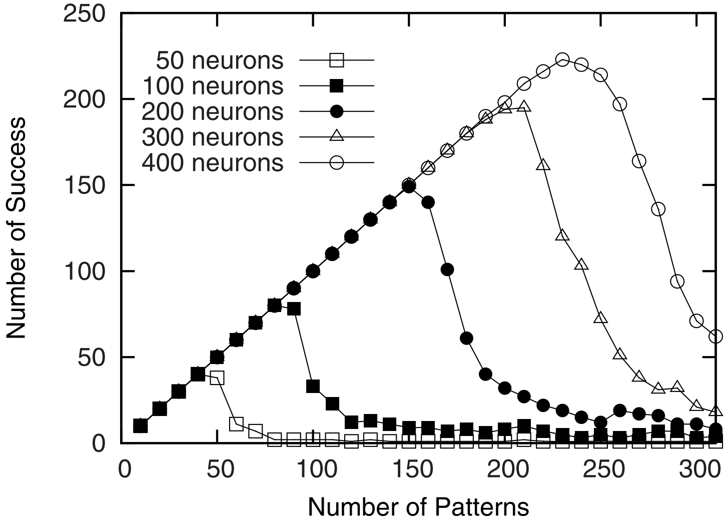


**Fig. 2.** Number of learned patterns ($\Delta w = 0.025$)

elements in a pattern are set to be −1 at first, then the half of the elements are chosen at random to turn to be 1.

The results of the simulations are shown in Fig. 1.

In Fig. 1, the horizontal axis is the number of patterns which are inputted to the network and the vertical axis is the number of patterns which are stored in the network successfully.

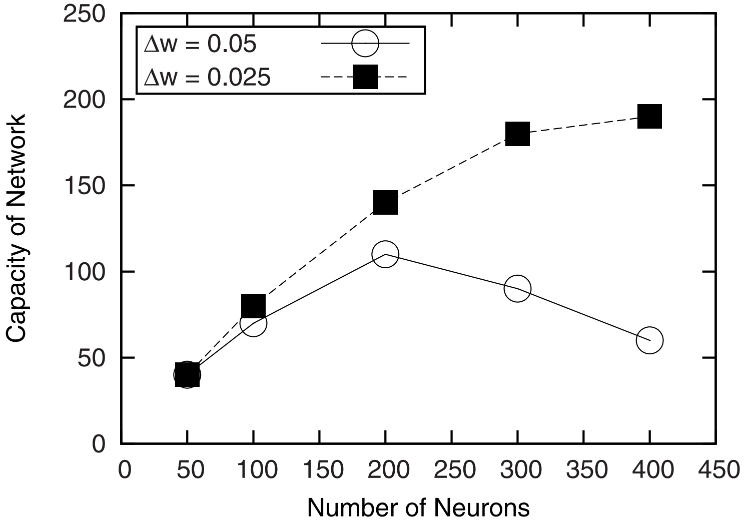


**Fig. 3.** Capacity of network

In the small numbers of input patterns corresponding to its size, the number of success is equal to the number of patterns. This means that the network has learned all of the patterns.

At some point—the value depends on its network size—the number of success comes to a peak. Over that point, the number of success decreases. This seems that too many patterns are conflicting to each other in the memory of the network, then almost no pattern can stay in the stable states.

Focusing on the differences between network sizes, the peak value moved to larger value along with the network size until 300 neurons. Although it is a natural thinking that the number of success grows up as the size increases, the number falls down from 300 to 400 neurons. This behavior may be due to the parameter $\Delta w$ which is the changing amount of the connection weights.

To verify this, the simulations were carried out with $\Delta w$ set to be 0.025. The results are shown in Fig.2,

Fig. 2 clearly shows that the peak value moved to larger value in large sizes and that the number of success increases as the number of neurons grows.

As described above, in the small numbers of input patterns, the network has learned all of the input patterns. We call the maximum number that satisfies this condition the "capacity of network" in this paper.

From Fig. 1 and Fig. 2, the capacities are picked up in every size of network and shown in Fig. 3 to summarize.

## 4 Capacity in Appropriate Parameters

In the preceding section, the capacity of network varies with the learning parameter $\Delta w$.

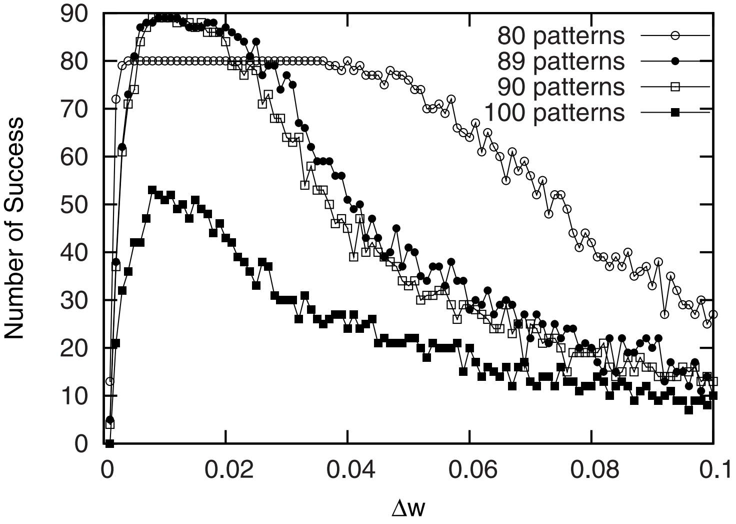


**Fig. 4.** Number of success with 100 neuron network

In this section, the simulations investigate the capacity after 100 sets of learning along with $\Delta w$ in the networks composed of 50, 100, 200, 300, or 400 neurons. In the simulations, we change $\Delta w$ from 0.001 to 0.1 in increments of 0.001.

The results of these simulations with the network composed of 100 neurons are shown in Fig. 4.

The horizontal axis is $\Delta w$ and the vertical axis is the number of success which is how many patterns the network learned. The key "80 patterns" means that the network received 80 patterns for input and the line shows how many patterns the network learned when 80 patterns are inputted.

From Fig. 4, all the 80 input patterns were learned within the range of $\Delta w$ from 0.004 to 0.036—"80 patterns" line reaches to 80—and so did the 89 patterns with the range from 0.009 to 0.012, but neither 90 nor 100 reached 90 or 100. In the case of "90 patterns", the line reached 89 not 90. Thus, the maximum capacity was figured out to be 89 with $\Delta w$ from 0.009 to 0.012.

In this way, we can find a maximum capacity at each size of network. Fig. 5 shows these maximum capacities with squares. For comparison, the capacities with the auto-correlative learning using the same patterns are also shown in Fig. 5 with circles. It should be restated that the capacity means the maximum number of stored patterns while the network can learn all the input patterns, in this paper.

Both of the capacities are seen to be proportional to the size of network, whereas the capacity of the incremental learning is above 11 times higher than that of the correlative learning.

In Fig. 6, $\Delta w$ which gives the maximum capacity is shown.

In this results, the appropriate $\Delta w$ is inverse proportional to the size of network.

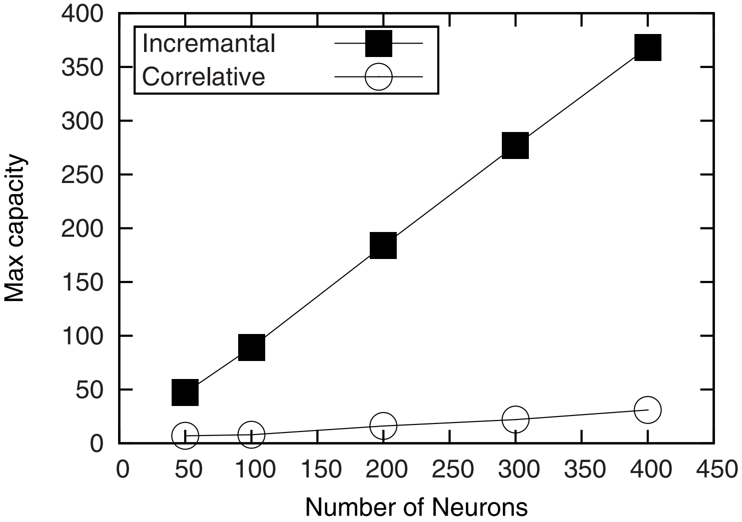


**Fig. 5.** Maximum capacity of network

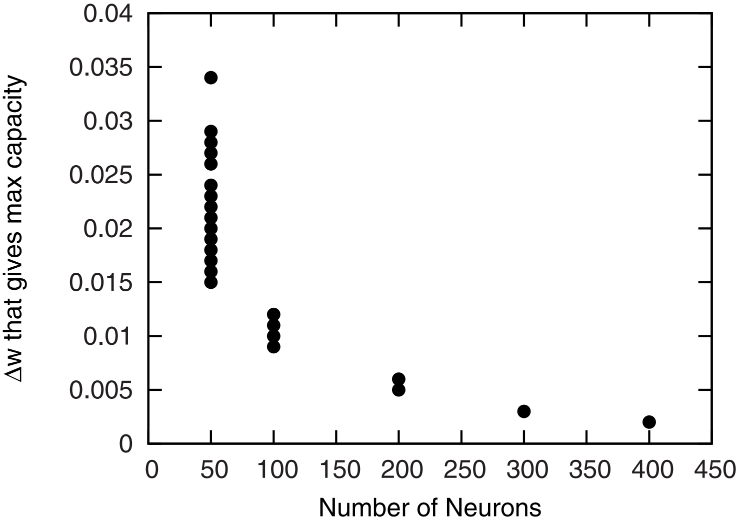


**Fig. 6.** $\Delta w$ which gives the maximum capacity

## 5 Conclusion

The capacity of the networks was investigated by changing the learning parameter. It turned out that the capacity of the network with the incremental learning increases in proportion to the size with appropriate parameter and that it is above 11 times larger than the one with correlative learning. The appropriate learning parameter is in inverse proportion to the size.

To investigate the effect of the length of the learning sets and the number of steps in a set is the future work.

## References

1. Asakawa, S., Deguchi, T., Ishii, N.: On-Demand Learning in Neural Network. In: Proc. of the ACIS 2nd Intl.Conf.on Software Engineering, Artificial Intelligence, Networking & Parallel/Distributed Computing, pp. 84–89 (2001)
2. Deguchi, T., Ishii, N.: On Refractory Parameter of Chaotic Neurons in Incremental Learning. In: Negoita, M.G., Howlett, R.J., Jain, L.C. (eds.) KES 2004. LNCS (LNAI), vol. 3214, pp. 103–109. Springer, Heidelberg (2004)
3. Watanabe, M., Aihara, K., Kondo, S.: Automatic learning in chaotic neural networks. In: Proc. of 1994 IEEE symposium on emerging technologies and factory automation, pp. 245–248 (1994)
4. Aihara, K., Tanabe, T., Toyoda, M.: Chaotic neural networks. Phys. Lett. A 144(6,7), 333–340 (1990)
5. Deguchi, T., Sakai, T., Ishii, N.: On storage capacity of chaotic neural networks with incremental learning. Memoirs of Gifu national college of technology 40, 59–62 (2005) (in Japanese)

# Improved Accuracy by Relearning and Combining Distance Functions

Naohiro Ishii, Takahiro Yamada, and Yongguang Bao

Aichi Institute of Technology,
Yachigusa , Yakusacho, Toyota, Japan
ishii@aitech.ac.jp

**Abstract.** The k-nearest neighbor(kNN) is improved by applying the distance functions with relearning and ensemble computations with the higher accuracy values. In this study, the proposed relearning and combining ensemble computations are an effective technique for improving accuracy. We develop a new approach to combine kNN classifier based on different distance functions with relearning and ensemble computations. The proposed combining algorithm shows higher generalization accuracy, compared to our previous studies and other conventional algorithms by artificial intelligence techniques. First, to improve classification accuracy, a relearning method with genetic algorithm is developed. Second, ensemble computations are followed by the relearning. Experiments have been conducted on some benchmark datasets from the UCI Machine Learning Repository.

**Keywords:** text classification**,** distance functions for classification**,** relearning, ensemble computation.

## 1 Introduction

Classification has been applied in many application fields, as the credit approval, pattern recognition, parts classification in industry and so on. Many inductive learning algorithms have been proposed for classification problems. For example, ID3, C4.5, k-Nearest Neighbor, Naïve-Bayes, IB, T2, Neural-Network, association rules et. Al. are developed. However, improving accuracy and performance of classifiers are still attractive to many researchers. In this paper, we focus on the k-nearest neighbor classification method(kNN)[1,2,3,4]. The kNN is a simple and effective method among instance-based learning algorithms. It is expected to provide good generalization accuracy by the kNN methods for a variety of real-world classification tasks as text classification and applications. Then, improving accuracy and performance of classifiers by the kNN, is still attractive to many researchers. In this paper, first, we present a new approach of relearning to improve the kNN classifiers based on distance functions with weights, which improve the performance of the k-nearest neighbor classifier. The weights of attributes of data, are optimized by applying genetic algorithm. Second, combining ensemble computation is developed as a final classifier, which is followed by relearning  mistaken classified data in the

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 926–933, 2008.

process. To improve the accuracy, not only the training data but also the testing data are important. The proposed method is developed from the 10-fold cross-validation method. Thus, it is shown that the relearning and combining procedures developed here, are effective to improve the classification accuracy based on the kNN. The proposed relearning and ensemble computation method shows a higher generalization accuracy, compared to other conventional learning algorithm.

## 2 Classification by Different Distance Functions

### 2.1 Distance Functions

The choice of distance function influences the bias of the k-nearest neighbor(kNN) classification. The most commonly used functions is the Euclidean Distance function (Euclid), which is defined as:

$$D(x, y) = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2}$$

where x and y are two input vectors (one typically being from a stored instance, and the other an input vector to be classified) and m is the number of input variables (attributes) in the application.

One way to handle applications with both continuous and nominal attributes is to use a heterogeneous distance function that uses different attribute distance functions on different kinds of attributes. The Heterogeneous Euclidean-Overlap Metric (HEOM) uses the overlap metric for nominal attributes and normalized Euclidean distance for linear attributes. This function defines the distance between two values x and y of a given attribute a as:

$$HEOM(x, y) = \sqrt{\sum_{a=1}^{m} d_a(x_a, y_a)^2}$$

where

$$d_a(x, y) = \begin{cases} 1, & if\; \mathrm{x}\; or\; y\; is\; unknown,\; else \\ overlap(x, y), & if\; a\; is\; nominal,\; else \\ \frac{|x-y|}{man_x - main_a}, & \end{cases}$$

and function overlap is defined as:

$$overlap(x, y) = \begin{cases} 0, \; x = y \\ 1, \; other \end{cases}$$

The Value Difference Metric (VDM), introduced by Stanfill and Waltz (1986)[4], is an appropriate distance function for nominal attributes. A simplified version of the VDM (without the weighting schemes) defines the distance between two values x and y of an attribute a as:

$$vdm_a(x,y) = \sum_{c=1}^{C} |\frac{N_{a,x,c}}{N_{a,x}} - \frac{N_{a,y,c}}{N_{a,y}}|^q = \sum_{c=1}^{C} |P_{a,x,c} - P_{a,y,c}|^q$$

where Na,x is the number of instances in the training set T that has value x for attribute a; Na,x,c is the number of instances in T that has value x for attribute a and output class c; C is the number of output classes in the problem domain; q is a constant, usually 1 or 2; and Pa,x,c is the conditional probability.

In [4], Wilson and Martinez proposed three new alternatives to overcome the weakness of VDM. The one is a Heterogeneous Value Difference Metric(HVDM) that uses Euclidean distance for linear attributes and VDM for nominal attributes. The other two distance functions are the Interpolated Value Difference Metric (IVDM). Wilson and Martinez also proposed a generic version of the VDM distance function, called the discretized value difference metric (DVDM).

### 2.2 Distance Functions for Optimization

The distance function in HEOM is defined in the previous section 2.1. To characterize the respective distance function for the training data, the weighted distance function is proposed in this paper as follows,

$$Weighted\ HEOM(x,y) = \sqrt{\sum_{i=1}^{n} \omega_i \times (x_i - y_i)^2}$$

where $\omega_i$ shows the weight of the i-th component of the data $x_i$ and $y_i$. Here, the weights are normalized as follows,

$$\sum_{i=1}^{n} \omega_i = 1, \quad \omega_i \geq 0$$

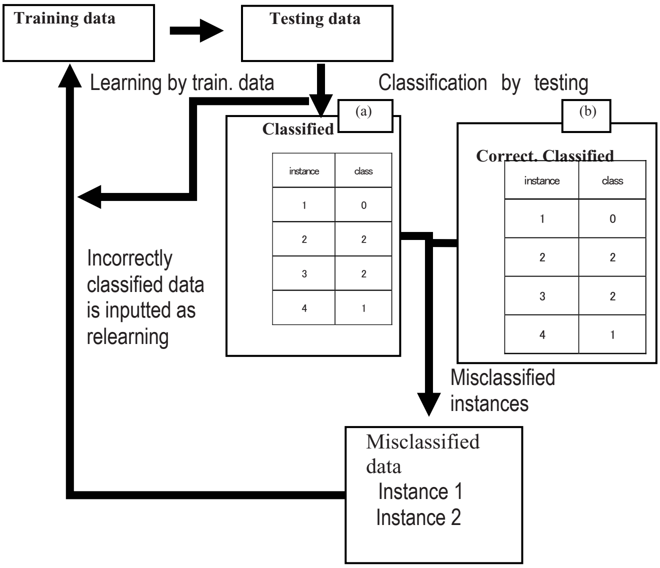
The problem, here is how to derive the optimized weights $\{\omega_i\}$. The optimized weights of the distance function, are computed by applying Genetic Algorithm(GA) to the training data. GA is a population-based iterative adaptive algorithm that uses selection, recombination and mutation operations based on natural selection and biological genetics.

## 3 Relearning Computation

To improve the classification accuracy, the similar data in the same class, will be important. The mistaken classified data will have a cue to improve the accuracy. Then, the mistaken classified data, is applied again in the relearning process, which is proposed in the following section.

### 3.1 Relearning Computation

In Fig.1, the first learning process is carried out by training data. The classification process is checked by using testing data in (a) in Fig.1. The classified testing data in (a) is compared with the correctly classified data table (b), which is given in advance. The



| instance | class |
|---|---|
| 1 | 0 |
| 2 | 2 |
| 3 | 2 |
| 4 | 1 |

| instance | class |
|---|---|
| 1 | 0 |
| 2 | 2 |
| 3 | 2 |
| 4 | 1 |

**Fig. 1.** Relearning process

**Table 1.** Experimental accuracy results without and with relearning

| DataSet | HEOM | | HVDM | | DVDM | | IVDM | |
|---|---|---|---|---|---|---|---|---|
| | Aver. | Re-learning | Aver. | Re-learning | Aver. | Re-learning | Aver. | Re-learning |
| Breast | 95.86 | **97.51** | 95.11 | **96.84** | 96.37 | **97.87** | 96.39 | **97.42** |
| Bridges | 58.39 | **72.07** | 60.49 | **72.16** | 60.09 | **81.14** | 58.43 | **77.54** |
| Flag | 56.32 | **70.27** | 58.83 | **82.11** | 55.30 | **80.37** | 54.09 | **81.03** |
| Glass | 75.33 | **79.57** | 71.76 | **85.86** | 58.66 | **66.19** | 77.27 | **89.95** |
| Heart | 80.04 | **87.63** | 80.35 | **87.95** | 82.36 | **88.52** | 80.20 | **87.11** |
| Heartlb | 86.56 | **92.00** | 72.70 | **84.71** | 85.55 | **90.05** | 85.90 | **90.59** |
| Heartswi | 95.40 | **97.77** | 96.81 | **97.00** | 96.18 | **97.85** | 95.66 | **98.49** |
| Hepatit | 81.83 | **90.05** | 79.12 | **89.43** | 80.66 | **91.35** | 86.69 | **90.66** |
| Promot | 80.36 | **90.54** | 88.11 | **97.61** | 88.97 | **97.63** | 88.69 | **95.52** |
| Average | **78.90** | **86.38** | **78.14** | **88.19** | **78.24** | **87.89** | **80.37** | **89.81** |

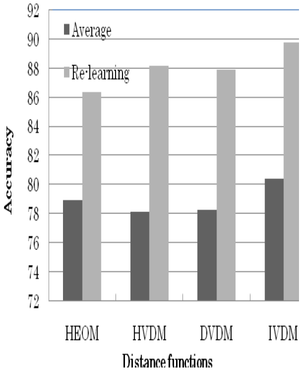


**Fig. 2.** Classification accuracy by relearning

misclassified data, instance 1 and instance 2, are applied in the second learning process as shown in Fig.1. In Table 1 and Fig.2, the experimental results of the classification accuracy without and with relearning process, are shown for the comparison.

### 3.2 Combining Ensemble Computation

To improve the accuracy, further in the classification, the operation of classifiers as a decision committee, is proposed. A committee as the final classifier, here, is composed of ensemble classifiers as committee members, each of which makes its own classifications that are combined to create a final classification result of the whole committee.

In this study, the 10-fold cross-validation method is applied, which implies the data is divided 10 subsets. Then, the 9 subsets are training data and the remaining one subset becomes the testing data. Thus, the ensemble computations are carried out for 10 times.

The values in the columns in Fig. 3(a), show the classes of instances by ensemble computations. The classes are summed and voted as shown in Fig. 3(b); 3 for the class 0, 7 for the class 1 and 0 for the class 2. Then, the class 1 is determined for the instance 1 as shown in Fig. 3(c). The ensemble computations with 10 times are combined in the classification of the 0 class with 3 times, and that of the 1 class with 7 times. Thus, the final classification becomes the 1 class by voting the majority of 7 times of 1. The class of the instance 1, is given as the class 1 in the data which is shown in the column, "correct" in Table 2, where 16 instances are assumed.

**Table 2.** Example of ensembleComputation

| instance | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 |
| 1 | 1.0 | 1.0 | 1.0 | 1.0 | 0.0 | 1.0 | 1.0 | 1.0 | 0.0 | 1.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| result | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| correct | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| error | | | | | X | | | | | X | X | | | | | |

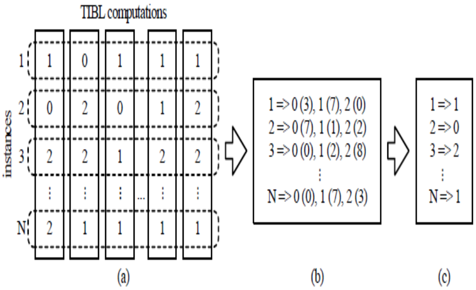


**Fig. 3.** Flow of computational algorithm

### 3.3 Experimental Results for Relearning and Combining Ensemble Computations

For evaluating the classification generalization accuracy of our algorithm with the relearning and combining ensemble computation, was tested on 7 benchmark dataset from the UCI Machine Learning Repository [8]. For the ensemble computations, we used the 10-fold cross-validation[1,4]. That is, the whole dataset is partitioned into ten subsets. Nine of the subsets are used as training set, and the 10th is used as the test set, and this process is repeated ten times, once for each subset being the test set. Then, classification accuracy is taken as the average of these ten runs. In our experiment, set parameters k = 3 and use four functions. In order to verify the combining ensemble and the combining functions, the following experiments are carried out. First, the combining ensemble processing is verified by comparing the method with relearning process and the method with relearning and combining ensemble computation by applying one distance function, HEOM, HVDM, DVDM and IVDM, respectively. Then, the generalization accuracy is shown in Table 3. The highest accuracy achieved for each dataset is shown in bold type as shown in Table 3. The method with the relearning and combining ensemble computation, is superior than only the relearning processing. The results in the distance function, IVDM, shows higher accuracy.

**Table 3.** Experimental results with relearning and ensemble computations

| DataSet | HEOM | | HVDM | | DVDM | | IVDM | |
|---|---|---|---|---|---|---|---|---|
| | Re-learning | Combining | Re-learning | Combining | Re-learning | Combining | Re-learning | Combining |
| Breast | 97.51 | **97.77** | 96.84 | **99.16** | 97.87 | **99.00** | 97.42 | **98.02** |
| Bridges | 72.07 | **76.00** | 72.16 | **82.89** | 81.14 | **91.81** | 77.54 | **92.61** |
| Flag | 70.27 | **71.62** | 82.11 | **96.41** | 80.37 | **95.57** | 81.03 | **95.41** |
| Glass | 79.57 | **83.18** | 85.86 | **95.28** | 66.19 | **80.85** | 89.95 | **94.55** |
| Heart | 87.63 | **89.45** | 87.95 | **97.02** | 88.52 | **95.87** | 87.11 | **94.28** |
| Heartlb | 92.00 | **93.09** | 84.71 | **88.67** | 90.05 | **93.30** | 90.59 | **96.65** |
| Heartswi | 97.77 | **97.78** | 97.00 | **98.18** | 97.85 | **99.09** | **98.49** | 98.18 |
| Hepati | 90.05 | **92.10** | 89.43 | **95.00** | 91.35 | **98.00** | 90.66 | **97.50** |
| Promo | 90.54 | **98.20** | **97.61** | 97.38 | 97.63 | **99.38** | **95.52** | 95.00 |
| Average | **86.38** | **88.80** | **88.19** | **94.44** | **87.89** | **94.76** | **89.81** | **95.80** |

The bold numerals in Table 3, show the higher accuracy value by the relearning process and ensemble computation. The distance function, IVDM shows higher accuracy value under the rough set condition with all instances[3, 11].

### 3.4 Combining Different Distance Functions for Classification

The algorithm for k-nearest neighbor classification from multiple different distance functions can be stated as: using simple voting, combining the outputs from multiple kNN classifiers. A method of combining the ensemble computations and the multiple distance functions without relearning, is shown in Fig.4.

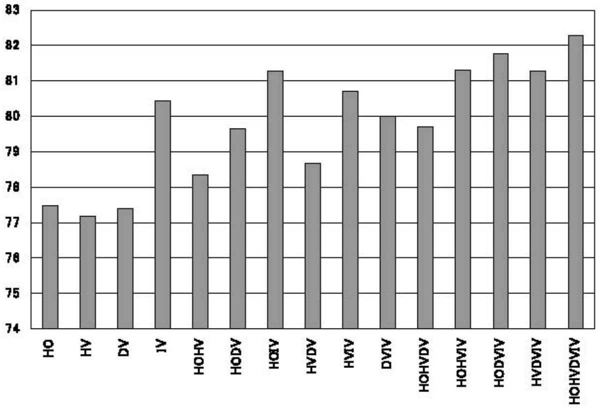


**Fig. 4.** Accuracy by combining functions without relearning

Fig.5 is computed under the relearning and the rough set condition of all, CP and CPP averages[3,11]. The combination of three distance functions, HVDVIV, shows the highest accuracy value in Fig.5, which also shows higher values than in Fig. 4.

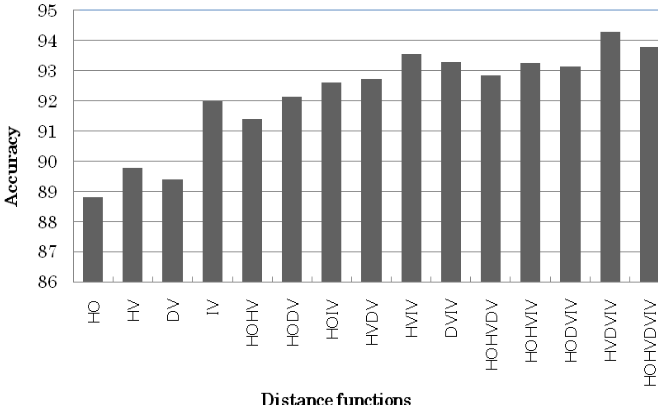


**Fig. 5.** Accuracy by combining functions with relearning

## 4 Comparison with Other Computational Methods

\Many inductive earning algorithms has been proposed for classification problems, such as ID3, C4.5, k-Nearest Neighbor, Naïve-Bayes, IB, T2, BP in neural net, association rules are developed as shown in Table 4. The bold in both relearning and combine method, in Table 4, show the higher values. The combine( combination of the relearning, the ensemble computations by using different distance functions under the average of All, CP and CPP[3,11]) shows the highest accuracy in Table 4.

**Table 4.** Comparison with Other Computational Methods

| Datasets | C4.5 | | | IB | | Bayes | BP | Re-learning | Combine |
|---|---|---|---|---|---|---|---|---|---|
| | Tree | P-Tree | Rule | IB1 | IB2 | | | | |
| Breast | 92.90 | 93.90 | 95.30 | 95.90 | 92.30 | 93.60 | 96.30 | **97.41** | **98.49** |
| Bridges | 68.00 | 65.30 | 59.50 | 53.80 | 45.60 | 66.10 | 67.60 | **75.73** | **85.83** |
| Flag | 59.20 | 61.30 | 60.70 | 63.80 | 59.80 | 52.50 | 58.20 | **78.45** | **89.75** |
| Glass | 68.30 | 68.80 | 68.60 | 70.00 | 66.80 | 71.80 | 68.70 | **80.39** | **88.46** |
| Heart | 73.30 | 72.10 | 80.00 | 76.20 | 68.90 | 75.60 | 82.60 | **87.80** | **94.15** |
| Hematite | 77.70 | 77.50 | 78.80 | 80.00 | 67.80 | 57.50 | 68.50 | **90.37** | **95.65** |
| Promo | 73.30 | 71.90 | 79.10 | 81.50 | 72.90 | 78.20 | 87.90 | **95.33** | **97.49** |
| Average | 72.97 | 72.56 | 74.01 | 74.53 | 68.47 | 71.36 | 75.59 | **88.07** | **93.45** |

## 5 Conclusions

Among the conventional developed classification algorithms, the instance-based learning algorithm using k-nearest neighbor is useful one. But, only the k-nearest neighbor algorithm has several weaknesses in the application. To cope with these problems, improvements for the basic k-nearest neighbor method, are needed in the classification problems. In this paper, a relearning and combining ensemble computations, is proposed by applying useful distance functions. The classification accuracy by the proposed method of relearning and ensemble computation, shows the superiority compared with other conventional methods.

Combining different distance approach will be useful for the further improvement in the classification accuracy. But, it is necessary to make clear what kind of combinations of different distance functions is better.

## References

[1] Wilson, D.R., Martinez, T.R.: An Integrated Instance-Based Learning Algorithm. Computer Intelligence 16(1), 1–28 (2000)

[2] Bao, Y., Tsuchiya, E., Ishii, N., Du, X.: Classification by Instance-Based Learning Algorithm. In: Gallagher, M., Hogan, J.P., Maire, F. (eds.) IDEAL 2005. LNCS, vol. 3578, pp. 133–140. Springer, Heidelberg (2005)

[3] Bao, Y., Ishii, N., Du, X.: A Tolerant Instance-Based Learning Algorithm. In: Dosch, W., Lee, R.Y., Wu, C. (eds.) SERA 2004. LNCS, vol. 3647, pp. 14–22. Springer, Heidelberg (2006)

[4] Wilson, D.R., Martinez, T.R.: Improved Heterogeneous Distance Functions. Journal of Artificial Intelligence Research 6, 3–21 (1997)

[5] Witten, I.H., Frank, E.: Data Mining Practical Learning Tools and Techniques. Morgan Kaufmann, USA (2005)

[6] Bay, S.D.: Nearest neighbor classification from multiple feature subsets. Intelligent Data Analysis 3, 191–209 (1999)

[7] Kaneko, S., Igarashi, S.: Combining Multiple k-Neighbor Classifiers Using Feature Combinations. IEICE Transactions on Information and Systems 1.2(3), 23–31 (2000)

[8] Merz, C.J., Murphy, P.M.: UCI Repository of Machine Learning Databases. Department of Information and Computer Science. University of California Irvine, Irvine (1998), `http://www.ics.uci.edu/~mlearn/MLRepository.html`

[9] Pawlak, Z.: "Rough Sets". Kluwer Academic Publishers, Dordrecht (1991)

[10] Pawlak, Z.: Decision Networks. In: Tsumoto, S., Słowiński, R., Komorowski, J., Grzymała-Busse, J.W. (eds.) RSCTC 2004. LNCS (LNAI), vol. 3066, pp. 1–7. Springer, Heidelberg (2004)

[11] Yamada, T., Yamashita, K., Ishii, N.: Text Classification by Combining Different Distance Functions with Weights. In: Proc. of SNPD 2006, pp. 85–90. IEEE Computer Society, Los Alamitos (2006)

# Influence of Character Type of Japanese Hiragana on Writer Recognition

Yoshinori Adachi, Masahiro Ozaki, and Yuji Iwahori

Chubu University, 1200 Matsumoto-Cho, Kasugai, Aichi, Japan 487-8501
ozaki@isc.chubu.ac.jp, adachiy@isc.chubu.ac.jp,
iwahori@ics.chubu.ac.jp

**Abstract.** The character used for the writer recognition was limited to 46 types of hiragana with a lot of curve parts, and the difference of the similarity values by the kind of character was examined. As a result, it has been understood that the similarity values were quite dependent on the type of character. Moreover, it turned out that the similarity value was greatly different according to the type of character. Moreover, it was confirmed that the recognition ratio was improved by using the steady character which writer could write similarly each time. In addition, it was confirmed that there exist the character that was appropriate for recognition even the similarity value was low (e.g. "く", "へ", "さ"), and the improper character even the similarity value was high (e.g. "け", "り", "し").

## 1 Introduction

Japanese is expressed by various characters, i.e. the Chinese character, the hiragana, the katakana, and the alphabet, etc. The hiragana is a type of character can be written from the adult to the child. It might be understood that the person with frequently use of the Chinese character is an educated person. However, there is movement to avoid the use of a difficult Chinese character so as not to cause the misunderstanding that comes from the misuse or miswrite of the Chinese character, and to write the hiragana as much as possible.

We have been studying the writer recognition for a long time [1-8]. In those studies, the Chinese character and the hiragana have been both used. And, the Chinese character showed that the recognition ratio was higher because Chinese characters have more strokes than hiragana. Moreover, it was confirmed that in the hiragana, there was a character not written stably because of the composition of the character, and a little use.

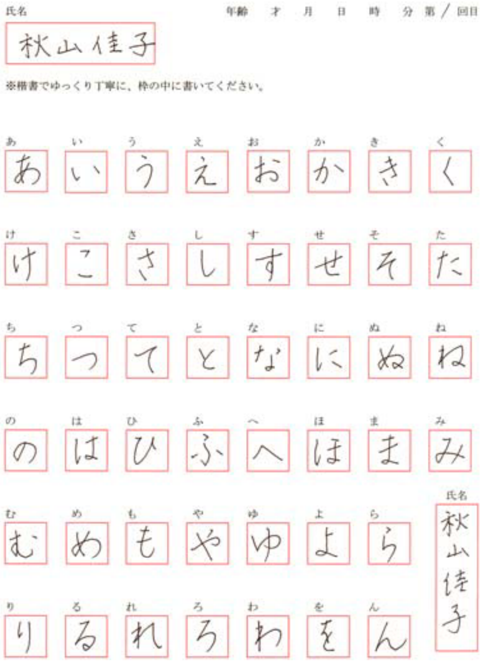
It is necessary to examine the feature of the hiragana about the writer recognition, because the hiragana can be used for writer recognition for all Japanese from young to elder. However, up to now, a feature of hiragana for the writer recognition was not well examined, and a systematic examination had not been done.

In this study, the influence of the type of hiragana on the writer recognition was examined for all hiragana (46 types of characters except "ゐ" and "ゑ" which are seldom used in ordinary writing), and the suitability to the writer recognition was examined.

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 934–941, 2008.

## 2 Recognition Experiment

To collect the specimen character, the special sheet was designed. From the previous research, the influence of the presence of frame on the writer recognition was examined and the framed character gave better recognition ratio. Therefore, in this study, we specified the frame to write down each character. Moreover, the suitable size of the character for writer recognition was also examined and 18mm square frame gave better results. In this study, the size of the frame was set to be 18mm square. Ten subjects (around 21-years-old) filled out one sheet a day, and they filled ten sheets. Figure 1 shows the example of the filled sheet.



氏名 年齢 才 月 日 時 分 第／回目

秋山佳子

※楷書でゆっくり丁寧に、枠の中に書いてください。

あ い う え お か き く

あ い う え お か き く

け こ さ し す せ そ た

け こ さ し す せ そ た

ち つ て と な に ぬ ね

ち つ て と な に ぬ ね

の は ひ ふ へ ほ ま み

の は ひ ふ へ ほ ま み

む め も や ゆ よ ら

む め も や ゆ よ ら

り る れ ろ わ を ん

り る れ ろ わ を ん

氏名

秋山佳子

**Fig. 1.** Example of filled sheet written by a subject

After the sheet image was input to a computer through a scanner (280dpi), each character was cut out one by one based on the each frame automatically. Therefore, collected characters became 4600 (10 subjects x 46 types x 10 times) characters in total.

The dictionary was formed by using five characters that had been written down in the first half. The similarity values were calculated for five characters that had been written down in the latter half. As a result, time series consideration also became possible.

The writer's feature was extracted by using a new local arc method. The chord length was adopted 13 dots which gave the highest accuracy in the previous study. The angle of the chord had been changed from 0° to 180° at every 15°. A curvature of the stroke of the character was obtained within the range from -5 to 5, and an appearance frequency of curvature was calculated as a feature vector of the character in 12x11 dimensions (12 angles x 11 curvatures).

The dictionary of each subject and each type of character was made as follows. First of all, the feature vector of five characters of each type of character was resolved to the eigenvalue and the eigenvector by the principal component analysis. The eigenvector was added so that the accumulation contribution rate of the eigenvalue might become 90% or more. The weighted average of the feature vector of five characters was calculated from the added eigenvector. In this study, the accumulation contribution rate of 90% or more was obtained by two eigenvalues. The similarity value was calculated by the commonly used inner product. The mean value of five characters was assumed to be a similarity value of each type of character, and it was obtained for each subject.

## 3 Experimental Results

Figure 2 shows the average similarity value of each type of character. Standard deviation is shown in the top of the bar with the line. There is a character with greatly different similarity values between those obtained from the own dictionary and the other's dictionaries (for example, “う”, “く”, “さ”, “へ”, “や”). On the other hand, there is a character with small deference between similarity values calculated from the own dictionary and the other's dictionary (for example, “あ”, “け”, “し”, “す”, “ね”, “も”, “ゆ”, “り”, “れ”, “わ”). As an example of these phenomena, Figure 3 shows the character change in time series consideration. Especially for “く”, it is understood that the shape of character change greatly. However, for “し” that is also one stroke character, a large similarity value is obtained from the own dictionary. “く” does have little curvatures like as “へ”. Therefore the similarity values of these characters are small.

The relation between the similarity values obtained from the own dictionary and the other's dictionary is depicted in Figure 4. Strong positive relation is obtained (r=0.846), indicating characters which have large similarity values are not necessarily suitable for writer recognition because even the similarity value obtained from the own dictionary is large, that obtained from he other's dictionary is also large.

It was examined whether there was a difference in the average similarity values by t-test. The results are shown in Table 1. The significance level 5% is indicated by * and 1% by **. A character of large t-value can be chosen from the table as a character that is appropriate for recognition because the similarity values obtained from the own dictionary and the other's dictionary is different. The characters of large t-value are chosen and shown in the figure in sequential order in Figure 5. The characters which have small t-values are listed in Table 2. The similarity values of these characters are not so small except “い” and “こ”. Especially the similarity value of “し”is 0.94 and it looks suitable for writer recognition, but it is a typical improper character (*t*=1.458, *p*>0.1 in t-test).

Next, the relation between the similarity value and t-value is shown in Figures 6 and 7. The similarity value obtained from the own dictionary hardly influenced with t-value and every similarity values are above 0.8 as shown in Figure 6 (correlation coefficient: r=0.221). The similarity value obtained from the others' dictionaries was scattered and no relation was observed as shown in Figure 7 (r=-0.156).

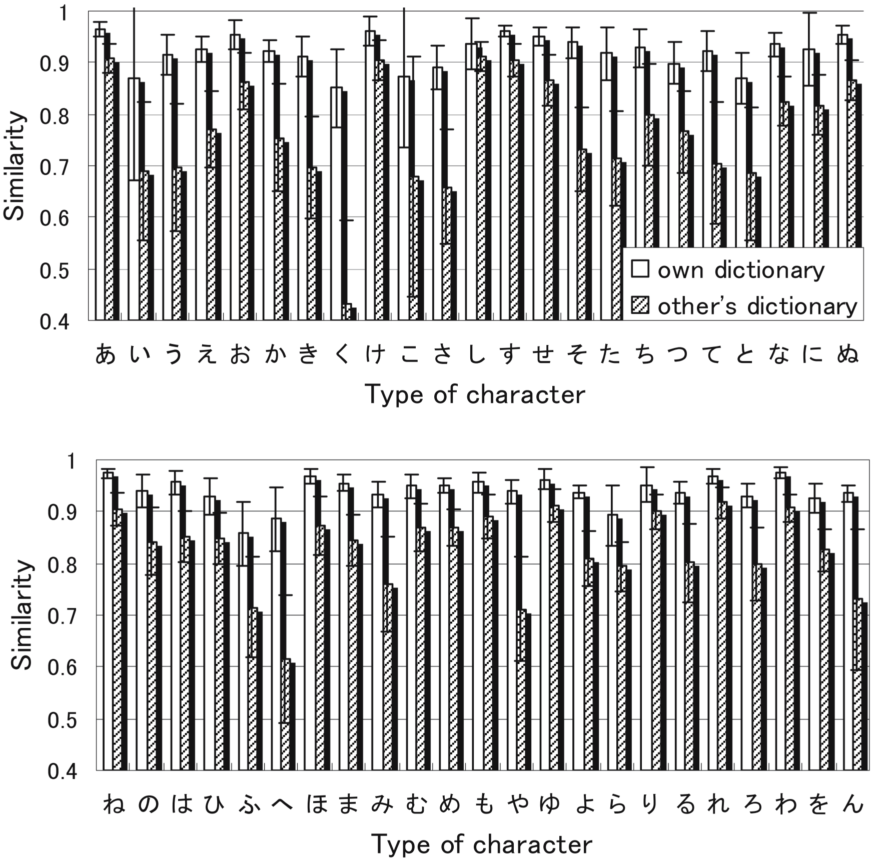


**Fig. 2.** Average similarity values of each type of character obtained from own dictionary and other's dictionary

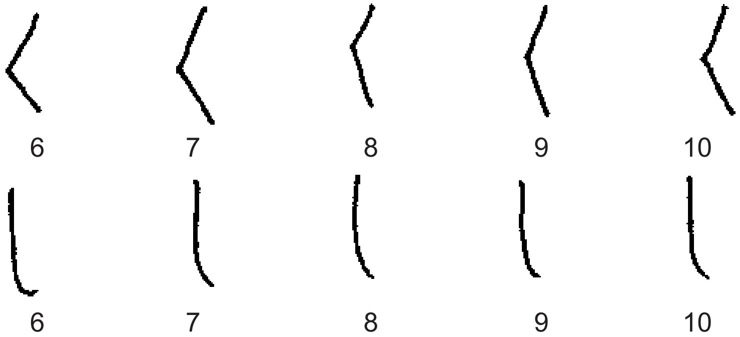


**Fig. 3.** Character shape change in time series

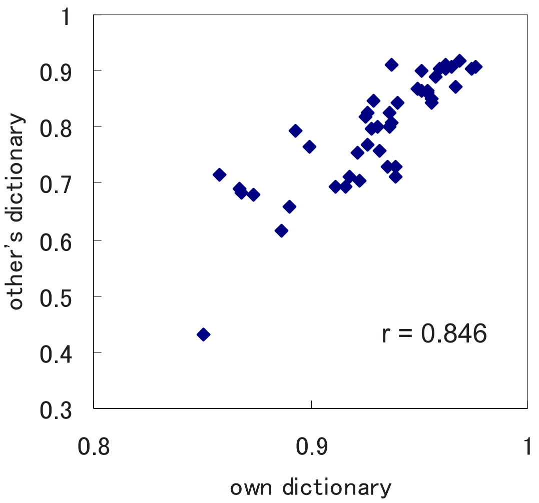


**Fig. 4.** Relation between similarity values obtained from the own dictionary and the other's dictionary

**Table 1.** Average similarity values and t-value

| No | Own dic | | Other's dic | | | | No | Own dic | | Other's dic | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Av | Stdv. | Av | Stdv | t | type | | Av | Stdv | Av | Stdv | t | type |
| 1 | 0.97 | 0.01 | 0.91 | 0.03 | 5.84 | あ** | 24 | 0.97 | 0.01 | 0.90 | 0.03 | 6.65 | ね** |
| 2 | 0.87 | 0.20 | 0.69 | 0.13 | 2.36 | い* | 25 | 0.94 | 0.03 | 0.84 | 0.07 | 4.23 | の** |
| 3 | 0.92 | 0.04 | 0.70 | 0.12 | 5.42 | う** | 26 | 0.96 | 0.02 | 0.85 | 0.05 | 6.10 | は** |
| 4 | 0.93 | 0.02 | 0.77 | 0.07 | 6.31 | え** | 27 | 0.93 | 0.04 | 0.85 | 0.05 | 4.21 | ひ** |
| 5 | 0.95 | 0.03 | 0.86 | 0.05 | 4.70 | お** | 28 | 0.86 | 0.06 | 0.72 | 0.10 | 3.97 | ふ** |
| 6 | 0.92 | 0.02 | 0.75 | 0.10 | 4.95 | か** | 29 | 0.89 | 0.06 | 0.62 | 0.12 | 6.23 | へ** |
| 7 | 0.91 | 0.04 | 0.70 | 0.10 | 6.43 | き** | 30 | 0.97 | 0.01 | 0.87 | 0.06 | 5.13 | ほ** |
| 8 | 0.85 | 0.08 | 0.43 | 0.16 | 7.37 | く** | 31 | 0.96 | 0.01 | 0.84 | 0.05 | 7.01 | ま** |
| 9 | 0.96 | 0.03 | 0.90 | 0.04 | 3.69 | け** | 32 | 0.93 | 0.02 | 0.76 | 0.09 | 5.70 | み** |
| 10 | 0.87 | 0.14 | 0.68 | 0.23 | 2.26 | こ* | 33 | 0.95 | 0.02 | 0.87 | 0.05 | 5.03 | む** |
| 11 | 0.89 | 0.04 | 0.66 | 0.11 | 6.08 | さ** | 34 | 0.95 | 0.01 | 0.87 | 0.04 | 6.56 | め** |
| 12 | 0.94 | 0.05 | 0.91 | 0.03 | 1.46 | し | 35 | 0.96 | 0.02 | 0.89 | 0.04 | 4.58 | も** |
| 13 | 0.96 | 0.01 | 0.90 | 0.03 | 5.61 | す** | 36 | 0.94 | 0.02 | 0.71 | 0.10 | 7.00 | や** |
| 14 | 0.95 | 0.02 | 0.87 | 0.05 | 5.19 | せ** | 37 | 0.96 | 0.02 | 0.91 | 0.03 | 4.28 | ゆ** |
| 15 | 0.94 | 0.03 | 0.73 | 0.08 | 7.56 | そ** | 38 | 0.94 | 0.01 | 0.81 | 0.05 | 7.52 | よ** |
| 16 | 0.92 | 0.05 | 0.71 | 0.09 | 6.17 | た** | 39 | 0.89 | 0.06 | 0.79 | 0.05 | 4.18 | ら** |
| 17 | 0.93 | 0.04 | 0.80 | 0.10 | 3.83 | ち** | 40 | 0.95 | 0.03 | 0.90 | 0.03 | 3.40 | り** |
| 18 | 0.90 | 0.04 | 0.77 | 0.08 | 4.68 | つ** | 41 | 0.94 | 0.02 | 0.80 | 0.08 | 5.40 | る** |
| 19 | 0.92 | 0.04 | 0.70 | 0.12 | 5.52 | て** | 42 | 0.97 | 0.01 | 0.92 | 0.03 | 5.04 | れ** |
| 20 | 0.87 | 0.05 | 0.68 | 0.13 | 4.21 | と** | 43 | 0.93 | 0.02 | 0.80 | 0.07 | 5.58 | ろ** |
| 21 | 0.94 | 0.02 | 0.82 | 0.05 | 6.72 | な* | 44 | 0.98 | 0.01 | 0.91 | 0.03 | 7.30 | わ** |
| 22 | 0.93 | 0.07 | 0.82 | 0.06 | 3.79 | に** | 45 | 0.93 | 0.03 | 0.83 | 0.04 | 6.52 | を** |
| 23 | 0.95 | 0.02 | 0.87 | 0.04 | 6.44 | ぬ** | 46 | 0.94 | 0.02 | 0.73 | 0.14 | 4.76 | ん** |

Av: average, Stdv: standard deviation, t: t-value, dic: dictionary.

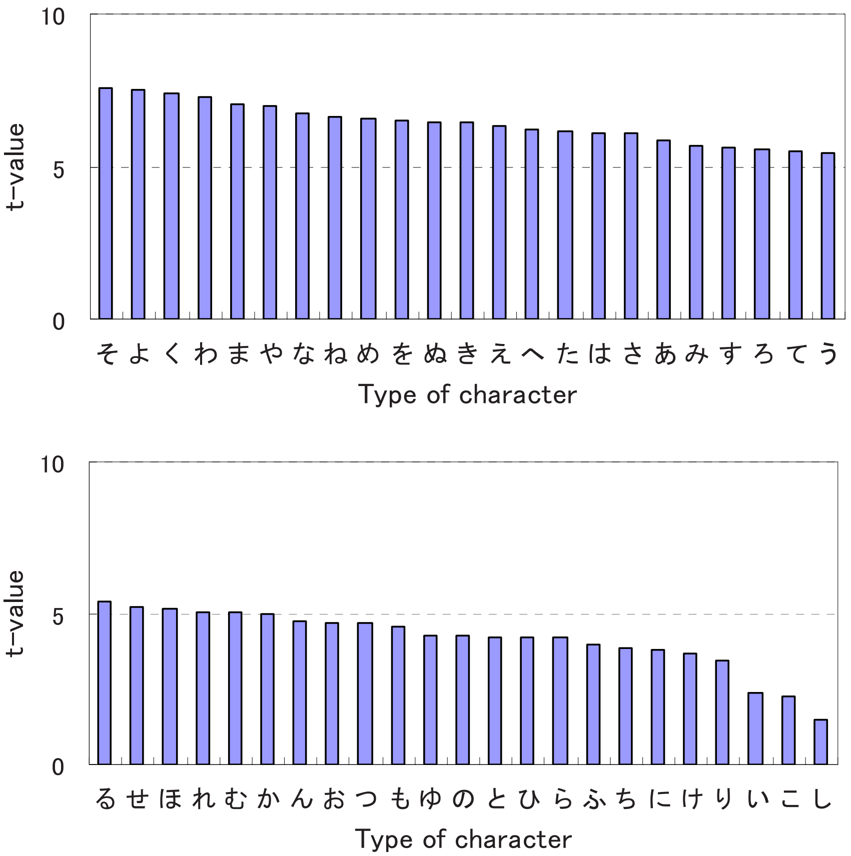


**Fig. 5.** t-values vs. character type

**Table 2.** Similarity values of small t-value characters

| | Own dictionary | | Other's dictionary | |
|---|---|---|---|---|
| | Av | Stdv | Av | Stdv |
| け | 0.96 | 0.03 | 0.90 | 0.04 |
| り | 0.95 | 0.03 | 0.90 | 0.03 |
| い | 0.87 | 0.20 | 0.69 | 0.13 |
| こ | 0.87 | 0.14 | 0.68 | 0.23 |
| し | 0.94 | 0.05 | 0.91 | 0.03 |

Av: average, Stdv: standard deviation.

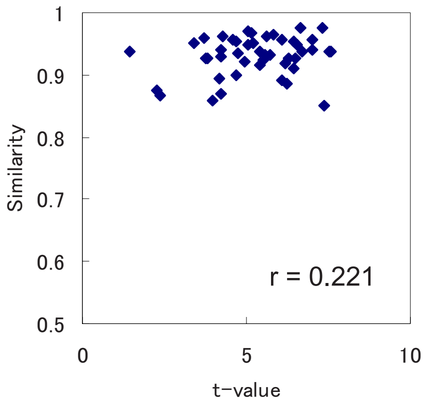


**Fig. 6.** Correlation between t-value and similarity value obtained from the own dictionary

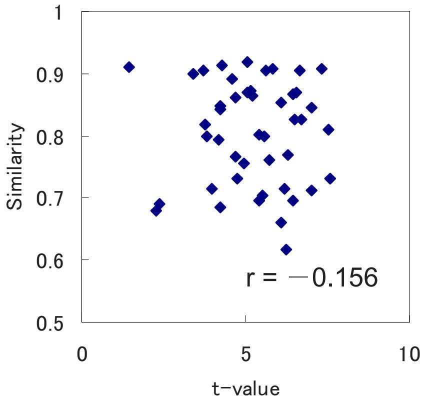


**Fig. 7.** Correlation between t-value and similarity value obtained from the other's dictionary.

## 4 Conclusion

From the above-mentioned results, we obtained the followings:

(1) The characters which have large similarity values are not necessarily suitable for the writer recognition.

(2) Most writers write characters which are changeable with time passage. Therefore own dictionary must be changed time to time.

(3) There are characters which express writer features well, even though the similarity values are small.

(4) Not to be similar to other's dictionary is more important than to have large similarity with own dictionary.

(5) The recommended characters for the writer recognition are "そ", "よ", "く", "わ", "や", "な", "ね", "め", "を" and not recommended characters are "け", "り", "い", "こ", "し".

In the present experiment, number of subject was only 10, and then the results might not be quite accurate. In the future, it will be necessary to increase the number of writers, and to examine the influence of the character type on the writer recognition further.

## References

[1] Ozaki, M., Adachi, Y., Ishii, N., Koyazu, T.: Fuzzy CAI System to Improve Hand Writing Skills by Using Sensuous. Trans. of IEICE J79-D-II(9), 1554–1561 (1996)

[2] Ozaki, M., Adachi, Y., Ishii, N.: Writer Recognition by means of Fuzzy Similarity Evaluation Function. In: Proc. KES 2000, pp. 287–291 (2000)

[3] Ozaki, M., Adachi, Y., Ishii, N.: Study of Accuracy Dependence of Writer Recognition on Number of Character. In: Proc. KES 2000, pp. 292–296 (2000)

[4] Ozaki, M., Adachi, Y., Ishii, N., Yoshimura, M.: Writer Recognition by means of Fuzzy Membership Function and Local Arcs. In: Proc. KES 2001, pp. 414–418 (2001)

[5] Ozaki, M., Adachi, Y., Ishii, N.: Development of Hybrid Type Writer Recognition System. In: Proc. KES 2002, pp. 765–769 (2002)

[6] Adachi, Y., Liu, M., Ozaki, M.: A New Similarity Evaluation Function for Writer Recognition of Chinese Character. In: Negoita, M.G., Howlett, R.J., Jain, L.C. (eds.) KES 2004. LNCS (LNAI), vol. 3214, pp. 71–76. Springer, Heidelberg (2004)

[7] Ozaki, M., Adachi, Y., Ishii, N.: Writer Recognition by Using New Searching Algorithm in New Local Arc Method. In: Khosla, R., Howlett, R.J., Jain, L.C. (eds.) KES 2005. LNCS (LNAI), vol. 3681, pp. 775–780. Springer, Heidelberg (2005)

[8] Adachi, Y., Ozaki, M., Iwahori, Y., Ishii, N.: Influence of presence of frame on writer recognition. In: Apolloni, B., Howlett, R.J., Jain, L. (eds.) KES 2007, Part II. LNCS (LNAI), vol. 4693, pp. 775–780. Springer, Heidelberg (2007)

# Learning Algorithm for Study Support in Web Study – Design of Prototype Model

Masahiro Ozaki[1], Yoshinori Adachi[1], and Naohiro Ishii[2]

[1] Chubu University, 1200 Matsumoto-Cho, Kasugai, Aichi, Japan 487-8501
ozaki@isc.chubu.ac.jp, adachiy@isc.chubu.ac.jp
[2] Aichi Institute of Technology, Yakusa-Cho, Toyota, Aichi, Japan 470-0392
ishii@aitech.ac.jp

**Abstract.** In the situation that the academic standards of Japanese university students have been falling down, how to force them to learn becomes very important. We have examined how to work the self-monitor strategy to improve their learning behavior. As a result, the condition of the self-monitor strategy was obtained: it must be small class, for the learner with a good grade, the self-monitor abbreviation is established, streaming is effective, and even in the large class the effect achieves by adopting the Web study. Furthermore, to improve the effect of Web study, the followings were found to be important: to specify the deadline of home study of quizzes, to present the evaluation of learning results by absolute grading, and to present the methods and techniques of instruction message at study.

## 1 Introduction

Recently, falling academic standards of the university student in Japan become problems. One of the causes is thought to be "Education with composure". However, the problem is not solved only by pointing out the cause. Therefore, many approaches on class improvement have been done by the university teacher, e.g. inventiveness of course content, development of learning material, usage of computer, and usage of Web teaching material, etc.

Under these circumstances, we develop the Web learning material including the use of the computer, and the Web education support system [1].

General Web study assumes the individual learner's ubiquitous study. However, neither a lot of Web learning materials nor the Web study support systems are maintaining the function to do the study support corresponding to a learning ability of individual student. For instance, study history information is gathered at study. However, because of the diversity of individual learner's understanding level, an effective study inference algorithm that makes the use of those information in real time has not been developed yet.

In the recent universities in Japan, there are many students who do not have the clear object after graduation. To fill credits for graduation, they are attending the class without a positive greediness for learning. Then, the dynamic Web learning material was developed in the previous study by judging the learner's understanding level

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 942–949, 2008.

dynamically [1]. However, the understanding level is greatly depending on the learner and the learning environment. An effective learning algorithm that carefully supports the learner has not been developed yet. Of course, those learning materials are not designed to support the learner who doesn't have the greediness for learning.

In the actual class, the learning outcome is measured by using the test result etc., but effective study support cannot be done based on it. It is necessary to consider learner's greediness for learning and learning behavior in each occasion. Then, we think roughly dividing the learning strategy into two strategies. First of all, it is necessary to judge what and how the learner voluntarily learns, and it is called a meta-recognition strategy. And, it needs the strategy to solve the study problem under the meta-recognition strategy, which is called a recognition strategy. We think that the learning outcome is obtained only after those two strategies action together effectively.

Our research purpose was to develop the meta-recognition strategy and the recognition strategy and also to develop the learning algorithm with the self-monitor strategy. And, an effective learning algorithm for the learner to achieve the study target will be developed based on these strategies.

In this paper, a study support system and Web learning material, which were necessary to develop the learning algorithm, were developed. And the results of the learning experiment by using them were analyzed to show the effectiveness of the learning algorithm.

## 2 Self-monitor

The self-monitor is a concept advocated by Snayder [2], he showed that the self-performed action and the idea became appropriate by self-controls based on the clue of the situation and the others' behavior. Ueki [3] described that the self-monitor was the present self-diagnosis of the understanding situation in the ongoing problem, and the intentional and voluntary use of this function was the self-monitor strategy. Kelley et al. [4] recorded the brain activity of the other-relation problem and the self-relation problem (activation of the self-knowledge) by fMRI. The case of the self-relation problem showed the strong activation in the frontal lobe. Moreover, Lane et al. [5] showed that similar occurred on the frontal lobe when the self-monitor was used. Then, it is considered that the self-monitor strategy is effective as the inferred learning strategy.

Nakagawa et al. [6] showed that didactics of the self-monitor evaluation improved inside motivation more than didactics of the attainment level self-evaluation in Japanese problem. Especially, they were shown to be effective in the study for subordinate position group. Here, the self-monitor evaluation training method is the one of the problem solving strategy, which monitors the execution process of the strategy, and does the self-management training of the evaluation and the error correction, etc., where the meaning of the skill were included in didactics. In addition, it is a method to self-evaluate the attainment level and execution process after trained to explain own method to others. Moreover, the self-evaluation of attainment level is a method to evaluate the target attainment level by comparing the objective targets to an external target which is set to be a criterion.

Ogura and Matsuda [7] reported that the self-evaluation condition raised motivation, while the external evaluation condition decreased motivation.

In addition, Shikauchi and Namiki [8], and Shikauchi [9] reported that they examined the effect of the evaluation structure on motivation of study, and the relative evaluation decreased the independent study behavior. The comparative assessment was an evaluation method which sets the standard from the group data, such as the deviation and five grade evaluation, etc.

These reports showed that the self-monitor strategy by absolute grading was an effective method for learner's motivation. However, whether the self-monitor strategy was used or not fully depended on the learner. In a word, if student attends the class only for the purpose of the acquisition of credits, the self-monitor strategy doesn't work effectively. It is shown that the self-monitor strategy as the meta-recognition strategy is difficult to establish.

Hayami [10] pointed out that the inner motivation of study gradually occurs from the external stimuli. In a word, it was suggested that the self-monitor strategy could be established by the repeated training.

In this research, we execute the learning experiment to train the self-monitor strategy, analyze the effect of didactics of the self-monitor strategy, and develop didactics that the self-monitor strategy works effectively. Furthermore, we aim to develop the learning algorithm for the self-study support.

## 3 Proposed Self-monitor Strategy

In the learning experiment by using the self-monitor strategy, it was general to train of executing the self-monitor strategy after the explanation. However, the same student attended two or more classes at the same time in the university. Therefore, even if the training of the self-monitor strategy was executed only in a specific class, it was a doubt whether it achieves the same effect of training. Then, we did not execute prior training of the self-monitor strategy, and executed methods and techniques of instruction that became the training of the self-monitor strategy by repeating in the learning experiment. In the experimental process, the self-monitor strategy were tried to be established naturally.

In this experiment, for the analysis of influence of the repetition of the self-monitor strategy was obtained, a brief explanation of the self-monitor strategy was given to the learner beforehand. In the study experiment, home study was obligated to reconfirm the understood content and to re-study the content not understood in the class by using the self-monitor strategy. To do so, the execution of quizzes that used 30 minutes of the latter half of the class, the self evaluation, and recording the study result in the confirmation vote (for submitting) and the confirmation sheet (for self management) were force to do. Moreover, study was directed to be advanced while obligating the description, submitting to the confirmation vote, and the record of the content studied by staying at home in the confirmation seat, and self-managing the content of study.

To train the self-monitor strategy by obligating self-manage, it was investigated whether to be able to make the learner established of the self-monitor strategy, and to obtain the study effect by the learning experiment. In addition, the learning algorithm to be effective for the self-monitor strategy was tried to be developed.

## 4 Learning Experiment by Self-monitor Strategy (1)

The learning experiments of the lecture course (about 100 students) and the maneuver class (about 20 students) were executed. The learning experiment in the lecture course was executed by the subject of the "information processing" class (half year) in the department that the author belonged and the maneuver subject was executed in an English class (full year) of another university. Moreover, the former was executed in two classes (the experiment group and the control group). The latter executed the learning experiment separately in three classes streamed by the result of the pretest.

Because the technique of the learning experiment was different like that, the same evaluation could be done. In each experiment, it was analyzed whether the self-monitor strategy was established.

Figure 1 is a result of the learning experiment in the lecture course (final exam). The horizontal axis shows the score and the vertical axis shows ratio (%). The average point of the experiment group was 50.4 and that of the control group was 49.7, and a significant difference was not admitted (from t-test: t=0.34, p>0.1). There was a possibility that the self-monitor strategy doesn't work effectively. It was thought that the learner was not intentionally using the self-monitor strategy.

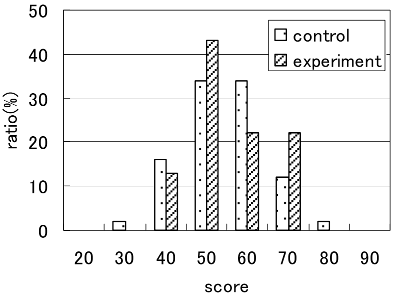


**Fig. 1.** Score of the end test of first term

The difference between the confirmation vote (every time, submit it immediately after execution of quizzes) that recorded the result of quizzes and the confirmation sheet (submit at the final class) was calculated. Then, the learner who recorded accurately (error 0) existed only 26.3%, and the error less than or equal 2 was about 52.6%. The learner who had five error or more existed as much as 9.6%. The learner of the more than half had the possibility of not adopting the self-monitor strategy.

Figure 2 shows the comparison of the results of the final exam (exclude the calculation problem) between the learner within 2 errors and the learner more than 3 errors in writing insertions. A horizontal axis shows the score and the vertical axis shows ratio (%). From the figure, the ratio in 40 points or less of the learner more than 3 errors in writing insertions (average: 45.8) is more than the learner within 2 errors (average: 53.3), and the score is also low. A significant difference is admitted as for the average point (from t-test: t=2.75, p≦ 0.01).

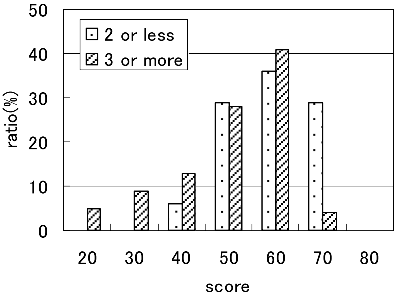


**Fig. 2.** Comparison of Score of final exam between within 2 error group and more than 3 error group

From these results, the learner with few errors in writing insertions was serious and seems to be adopting the self-monitor strategy. However, the greater parts of students did not adopt the self-monitor strategy in this experiment. The result of the self-monitor strategy was not able to be measured enough.

Then, the same study experiment was comparatively tried about the maneuver classes of the few people (about 20-40 students in each class). However, because the learning experiment was executed in an actual class, it was not possible to divide into the experimental group and the control group as in the lecture class experiment. Moreover, the streaming shown in Table 1 was adopted in an English maneuver class, and the learning experiment was executed here under those class organizations. The contents of the study and of the test were made division into three classes (A, B, and C) from the result of the pretest as shown in Table 1 according to the ability as shown in the table. A basic problem set the content studied in the class, and the applied problem set the content not studied in the class.

**Table 1.** Streaming class

| Class | Level | Number of quizzes | |
|---|---|---|---|
| | | Basic | Application |
| A | second STEP class or more | 10 | 10 |
| B | second semi-STEP class | 15 | 5 |
| C | third STEP class | 20 | 0 |

Figure 3 shows the average score of each test in three classes. The horizontal axis shows the type of test and the vertical axis shows the score. The average score of class A (69.7) is the highest at the pretest. The average scores of three classes rise in the test result at the end of the first term. The average scores are 77.1, 69.1, and 55.3 respectively. The average score of class C is the lowest, and the growth rate of class B is highest. However, because the difficulty levels of the examination of class A is the highest, it can be said that the growth rate of class A is also high. Each average score is 72.2, 67.5, 50.4, and the average score of class A is raised most in the tests at the end of latter term even though half the number of quizzes is the applied problem.

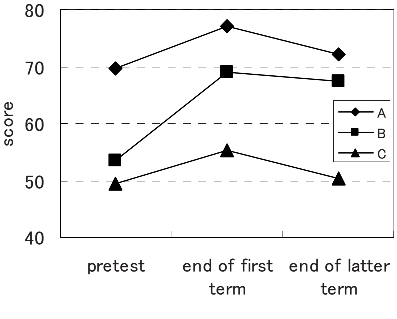


**Fig. 3.** Change of average score

The average agreement ratio between declared error number and actual error number at each class was calculated. Class A's ratio was higher than those of classes B and C. From the result, the learner in class A was regarded to have adopted the self-monitor strategy as O'Malley and Chamet [11] pointed out. Moreover, because of the comparatively few students in the maneuver class and possible of professor's individual counseling to the learner, it was thought that there was little error in writing insertion compared with the learning experiment in the lecture class.

However, it was a rashness to think that the self-monitor strategy was established from this learning experiment. Then, the learning experiment tried again in the lecture course where the self-monitor strategy was not adopted.

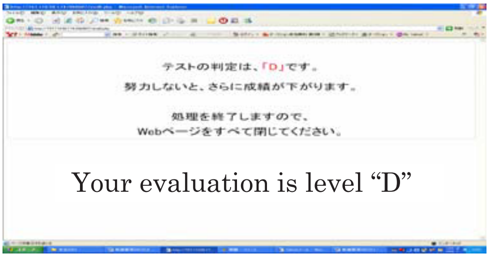
## 5 Learning Experiment by Self-monitor Strategy (2)

Here, to judge whether the self-monitor strategy was established by add the streaming and the individual counseling, the learning experiment was executed again in the large lecture class. However, it was difficult to add the streaming and the individual counseling like as a maneuver class. It was difficult to execute the streaming in an actual class. Therefore, the result of quizzes decided to be presented to the learner in every time through Web as shown in Figure 4 as the class composition shown in Table 2.

**Table 2.** Score distribution according to ability

| Class | A | B | C | D |
|---|---|---|---|---|
| Score | 80-100 | 61-79 | 51-60 | 0-50 |

Figure 4 is an ability judgment displayed on the Web page immediately after quizzes are executed, and, as a result, the learner knows the present level. These evaluations used the absolute grading for keeping motivation high. In addition, to inform of learner's current state, the message for methods and techniques of instruction also affixed it. And, the deadline of handing in the grading result of quizzes was clarified. The problem of answered wrong was directed the learner to study again by next week.



テストの判定は、「D」です。
努力しないと、さらに成績が下がります。

処理を終了しますので、
Webページをすべて閉じてください。

Your evaluation is level “D”

**Fig. 4.** Classification screen

The quizzes on Web were the empty column replenishment problem same as the previous experiment and the choices showed on a right page on the screen. Moreover, the correct answer was displayed in blue, and the wrong answer was displayed in red, and an empty column was displayed in purple. In addition, the confirmation vote was presented at the position of the choices (right page) for the sake of the self-monitor. After the grading was confirmed, the number of correct answers, the number of wrong answers, the number of understanding, and the number of not understanding were forced to input at the input column to evaluate the study result by one’s self.

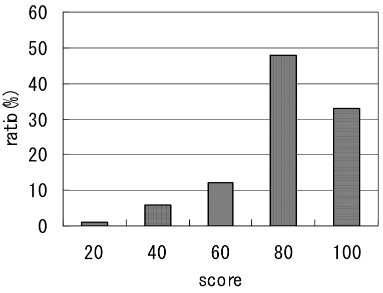


**Fig. 5.** Score of final exam

Figure 5 shows the score distribution of the final exam by the Web learning experiment. A horizontal axis shows the score and the vertical axis shows the ratio. Over all, scores are high, and the average score is 81.48. The average score of the final exam of the previous experiment with same problems was 50.4 points, therefore 81.48 is considerably high score.

It is possible to think that this learning experiment is effective for the improvement of the study result because there is no big difference in learner's ability between these classes.

## 6 Conclusion

From the result of the leaning experiment, it is understood that learner's greediness for learning is very important to obtain an effective learning outcome. The condition that the self-monitor strategy worked effectively was analyzed from the experimental results and the followings were obtained;

(1) For good understanding, it must be small class.
(2) For the learner with a good study grade, the self-monitor strategy might be established.
(3) Streaming is effective for good understanding.
(4) Even the large class, the Web study gives good effect for self-monitor study.

It has been understood that the Web study is effective though it is not possible to conclude from a little learning experiment. Then, the effect of the Web learning experiment that had been executed for the large class was analyzed. As a result, the followings are found to be important,

(1) Specify the deadline of home study of quizzes by using Web,
(2) Present the evaluation of learning results by absolute grading, and
(3) Present the methods and techniques of instruction message at study.

As for effective "Learning algorithm" that achieved effective methods and techniques of instruction, it was confirmed that the meta-recognition strategy like the learner's self-monitor had to synchronize with an individual learning strategy from these results. Moreover, it can be thought that it is greatly influenced also from the learning environment from the experiment on Web.

It will be necessary to develop "Learning algorithm" that operates after repeating the study experiment with Web teaching material, and established of the self-monitor strategy in the future.

## References

[1] Ozaki, M., Koyama, K., Takeoka, S., Adachi, Y.: Development of teaching materials which dynamically change in learning process. In: Negoita, M.G., Howlett, R.J., Jain, L.C. (eds.) KES 2004. LNCS (LNAI), vol. 3214, pp. 77–82. Springer, Heidelberg (2004)

[2] Furuhata, K. (ed.): Cyclopedia of social psychology, Yuhikaku (1994) (in Japanese)

[3] Ueki, R.: What guidance is necessary for established of the self-monitor strategy. Research in educational psychology 52, 277–286 (2004)

[4] Kelly, et al.: Finder the self?: an event-related fMRI study. J. of Cognitive Neuroscience 14, 785–794 (2002)

[5] Lane, et al.: Neural activation during selective attention to subjective emotional responses. Neuroreport 8, 3969–3972 (1997)

[6] Nakagawa, M., Moriya, T.: Effect of didactics on unit study of national language. Research in educational psychology 50, 81–91 (2002)

[7] Kokura, Y., Matsuda, F.: Effect of evaluation on motivation in student. Research in educational psychology 36, 144–151 (1988)

[8] Shikauchi, M., Namiki, H.: Effect of evaluation structure on motivation of study in child. Research in educational psychology 41, 36–45 (1990)

[9] Shikauchi, M.: Effect of evaluation of achievement on motivation in child. Research in educational psychology 41, 367–377 (1993)

[10] Hayami, T.: Achievement motivation located between inside drive and outside drive. Psychology commentary 38(2), 171–193 (1995)

[11] O'Malley, J.M., Chamot, A.U.: Learning strategies in second language Acquisition. Cambridge University Press, Cambridge (1990)

# Municipal Solid Waste Site Location Using a Fuzzy Logic Approach

Beatriz S.L.P. de Lima[1,2], Maria C.M. Alves[2], Alexandre G. Evsukoff[1], and Ian N. Vieira[1]

[1] COPPE and [2] Polytechnic School /Federal University of Rio de Janeiro
P.O. Box 68506, 21941-972 - Rio de Janeiro, Brazil
bia@coc.ufrj.br, alves.mcm@gmail.com, evsukoff@coc.ufrj.br, ianvieira@gmail.com

**Abstract.** This work presents a decision-support system (DSS) based on fuzzy logic to select the best site for the installation of a Municipal Solid Waste (MSW) landfill. Selection of a new site requires the evaluation of many criteria involving environmental, social and economic data. Such a large range of information comprises not only quantitative, but also qualitative criteria. In order to deal with this peculiarity, the developed system used fuzzy rules due to its ability to deal with linguistic variables and most of human reasoning. Conventional approaches tend to be less effective in dealing with the imprecise or vague nature of the linguistic assessment. A case study using real data of a recent MSW location in Brazil is presented where the developed system showed good results.

**Keywords:** Fuzzy decision-support system, waste management, solid waste landfill.

## 1 Introduction

Final disposal of solid waste in most developing countries has been just a matter of transporting the collected waste to the nearest available open space and discharging it without any special care, leading to the so-called dumpsites. In Brazil, a great governmental effort is on course in order to minimize inadequate MSW disposal. Nevertheless, the selection of new landfills sites is not a straightforward task. Many aspects, like environmental features, social impact assessment, and cost considerations must be accounted for in order to point to an adequate management of MSW. Some of those aspects deal with linguistic terms that involve vagueness and imprecision which are not simply to put into numbers and to be modeled.

Knowledge-based systems applied to geotechnical engineering started to be discussed in the nineties [1]. Decision support analyses are presented in many environmental issues [2,3] including decision models with the purpose of assisting designers in managing waste disposal and treatment [4,5].

In this work, a computational tool was developed based on a fuzzy decision approach that is able to integrate qualitative and quantitative information and to

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 950–957, 2008.

incorporate the vagueness and imprecision due to this type of decision-making process. Some of the above mentioned references do not deal with qualitative information, for example socio-economic issues, that may be critical to the landfill site selection. The developed computational system involves a DSS based on a fuzzy model, which is applied to a real case study in the State of Rio de Janeiro, in the Southeast region of Brazil.

## 2 Landfill Overview and Brazilian Scenario

Brazil is composed of five geographic regions, which are quite different with regard to their size, economic development and demographic density. The urban concentration of the population is around 80% and the amount of collected urban waste in Brazil is approximately 230,000 t/day.

There is not, in nature, a place that is considered ideal for landfill implantation. Nevertheless, the more careful the evaluation of the available areas for implantation is, the smaller the risk of turning it into an environmental problem and the lower the implantation and operation costs will be. The most adequate area will be the one that best fits all the pertinent criteria, either technical, environmental, social, economic, or even legal. The more important environmental points to be aware of are related to the geological, geotechnical and hydro geological conditions, protection of water bodies, and meteorological conditions. The Brazilian technical norm for design criteria, implantation and operation of a MSW landfill establishes many recommendations, the most important ones deal with limit values for: permeability of the foundation soil, depth of ground water level, distance to the nearest surface water body, and ground slope. A number of other conditions that are not so easily put into numbers must be evaluated, as for example, the conditions of assessment routes, proper use of the soil, occupation of access routes, acceptance of local communities, cost of the land, availability of cover material, political issues and so on. Additionally, some operational data, such as distance to residential areas, distance to the collect centre, and good conditions of accesses for heavy trucks must be taken into account.

The computational tool developed in this work deals with a Multi-Criteria Decision Making system, which will be discussed in detail in next section. This system involves some conditions mentioned above and many other decision criteria that must be considered when selecting a landfill site. In this paper, those criteria are divided into three groups, as shown below.

The first group deals with ***environmental criteria*** as described below. The Brazilian Norm requires the minimum values for some of those criteria.

**Distance to surface water bodies:** The landfill sites cannot be settled in a distance less than 200m from any water body.

**Soil Permeability:** must be less than $5 \times 10^{-5}$ cm/s to avoid contamination by landfill leachate.

**Depth to the ground water level:** The distance between the landfill bottom surface and the highest level of the ground water must be at least 1.5 m.

**Distance from airports:** greater than 3 km from any airport.

**Extension of drainage basin:** The extension of the catchment's area must be as small as possible in order to avoid large water volumes in the landfill site.

**Land use:** The site must be out of environmentally protected areas and should be in industrial or agricultural areas.

Four ***social criteria*** compose the second group, as listed below. The last three criteria are considered as qualitative criteria.

**Distance from residential areas:** minimum distance of 1000m from residential areas.

**Distance from low-income communities:** Waste deposit sites can attract low-income and unemployed people in a attempt to take their living out from landfill. This fact can cause a serious social problem for municipalities that need to implement sustainable mechanisms for employment of these people.

**Occupation of access routes:** It is desirable for truck traffic to occur within areas with low demographic density.

**Problems with local communities:** The acceptance level of the surrounding communities must at least be satisfactory.

The third and last group is composed of ***economic criteria*** that are mostly composed of qualitative criteria (except for the lifetime and ground slope)**:**

**Availability of cover material:** It is recommended that a suitable amount of cover material is available near the landfill site.

**Lifetime**: The minimum lifetime should be at least 10 years.

**Land cost: A** good negotiation for land use is always desirable

**Investment in infrastructure:** Complete infrastructure is desirable so that it will minimize installation costs.

**Access to heavy trucks:** Good pavement roads without hard ramps and curves are the best conditions for heavy trucks traffic.

**Distance from the collect center:** It is expected that distances should be as small as possible from the collect center to reduce costs.

**Ground slope** must be between 1% and 30%. The local morphology is supposed to facilitate the leachate collection system for the treatment before effluent discharge into the water bodies.

Those three groups were implemented separately in the system. In this way, the decision maker can access not only the global grade but also the grades for each particular group. The global grade is the weighted average of the three groups, so different weights can be furnished for each group according to the decision maker judgment.

The criteria described above represents quantitative and qualitative information and some of the measures may be corrupted by uncertainty, such that the decision problem is difficult to be modeled using classical multi-criteria decision methods. In this work, a fuzzy decision approach is used in order to integrate qualitative and quantitative information, to incorporate the uncertainty and to sweep off the discontinuity from decision process as much as possible, as described in next section.

## 3 The Fuzzy Decision Approach

In the fuzzy system's approach for Multi-Criteria Decision Making (MCDM), each decision is considered as a fuzzy set defined in the domain fixed by the criteria. This section describes the representation of a fuzzy rule-based MCDM approach, which is based on the fuzzy pattern matching approach.

Consider the criteria as input variables and the possible decisions represented by the set $\mathbf{D} = \{D_1 \ldots D_m\}$. The solution to the decision problem is to assign a decision $D_k \in \mathbf{D}$ corresponding to an observation set of the criteria. In a general application, the input variables' values may be numeric (discrete or continuous) or nominal. Fuzzy sets allow a unified representation for nominal and numeric variables as fuzzy sets.

Each input variable $x_i$ can be described using ordered linguistic terms in a *descriptor set* $\mathbf{A}_i = \{A_{i1}, \ldots, A_{in_i}\}$. When the variable is nominal, the descriptor set is the set of possible values for the variable (or a combination of them). When the variable is numeric, the *meaning* of each term $A_{ij} \in \mathbf{A}_i$ is given by a fuzzy set defined on the variable domain. The process of computing such fuzzy sets is known as the fuzzification of the variable. Fuzzification is an important issue since it provides the linguistic-to-numeric interface that allows dealing with variable values as linguistic terms.

In order to simplify computations, fuzzy sets are computed by strong fuzzy partitions of the input variables domain, such that:

$$\forall x_i(t), \quad \sum_{j=1..n_i} \mu_{A_{ij}}(x_i(t)) = 1 \ . \tag{1}$$

An easy way to parameterize such fuzzy partitions is to use triangular membership functions that are completely determined by the centers of triangles, which may be considered as prototype values for the corresponding fuzzy sets. For each input variable (criterion), the result of the fuzzification is the fuzzification vector that generalizes the information contained in the input variable. The fuzzification vector is computed in the same way if the variable is numeric or nominal as:

$$\mathbf{u}_i(t) = \left(\mu_{A_{i1}}(x_i(t)) \ldots \mu_{A_{in_i}}(x_i(t))\right) \ . \tag{2}$$

where $\mu_{A_{ij}}(x_i(t))$ is the fuzzy membership function of the variable to the fuzzy set $A_{ij}$. If the variable is nominal then the fuzzification vector is a binary vector, with a unitary membership corresponding to the observed nominal value and zero membership to all other values.

A fuzzy rule relates input linguistic terms $A_{ij} \in \mathbf{A}_i$ to the decisions $D_k \in \mathbf{D}$ in rules written as:

$$\textit{if } \mathbf{z}_i(t) \textit{ is } A_{ij} \textit{ then the decision is } D_k \ . \tag{3}$$

where $\mathbf{z}_i(t) \in X^q$, $q < p$, represents a subset of input variables that are considered in the multidimensional rule and $A_{ij}$ is a fuzzy set in the multi-variable fuzzy partition defined over $X^q$.

For applications with a large number of variables, reasonable results can be achieved using partial output aggregation of mono-variables sub models [8].

In this work, the confidence factor is $\varphi^{i}_{jk}$. A set of rules (or a rule base) for each input variable defines a sub-model that can be represented by the matrix $\mathbf{\Phi}_i$ where the lines $j = 1...n_i$ are related to the terms in the input variable descriptor set $\mathbf{A}_i$ and the columns $k = 1...m$ are related to decisions in the set $\mathbf{D}$. The values $\varphi^{i}_{jk} \in \{0,1\}$ represent the rules linking the term $A_{ij} \in \mathbf{A}_i$ to the decision $D_k \in \mathbf{D}$, such that a value $\varphi^{i}_{jk} = 1$ means that the observation of the term $A_{ij}$ is related with the decision $D_k$ and corresponding rule is present in the model.

The rule base is the kernel of the fuzzy decision model. The fuzzy inference model is consistent if the confidence factor represented by values $\varphi^{i}_{jk} \in [0,1]$ are used in the rule base, as is often the case in classification problems. Nevertheless, factionary rule weights are difficult to assign by human experts and only binary rule weights were used in this work. The fuzzy inference computes partial outputs for each sub-model from the input variables values. There is one fuzzification vector for each sub-model, but a sub-model can be related to more than one input variable. The choice of which variables must be considered in a multi-variable sub model is application dependant.

The output of the fuzzy system is the decision membership vector (or the fuzzy model output vector) $\mathbf{y}(t) = (\mu_{D_1}(\mathbf{x}(t)),\ldots,\mu_{D_m}(\mathbf{x}(t)))$, where each component $\mu_{D_k}(\mathbf{x}(t))$ is computed from an $\mathbf{x}(t)$ in two steps. The first one computes a partial output $\mathbf{y}_i(t)$ for each sub model, and then the final output $\mathbf{y}(t)$ is computed by aggregation of the partial outputs. The partial output is the vector $\mathbf{y}_i(t) = (\mu_{D_1}(x_i(t)),\ldots,\mu_{D_m}(x_i(t)))$, of which the components represent the membership value of the possible decisions considering only the information in the sub model $i$.

Adopting strong normalized fuzzy partitions and using the sum-product operator for the fuzzy inference, the partial output is computed from the membership vector $\mathbf{u}_i(t)$ and the rule base weights' matrix $\mathbf{\Phi}_i$ by the max-min fuzzy composition operation as:

$$\mu_{Dk}(x_i(t)) = \max_{j=1...n_i}\left(\min\left(\mu_{A_{ij}}(x_i(t)), \mathbf{\Phi}_i(A_{ij}, D_k)\right)\right) . \tag{4}$$

The final output is computed by the aggregation of all partial conclusions $\mathbf{y}_i(t)$ by an aggregation operator $\mathbf{H} : [0,1]^p \to [0,1]$ as:

$$\mu_{D_k}(\mathbf{x}(t)) = \mathbf{H}\left(\mu_{D_k}(x_1(t)),\ldots,\mu_{D_k}(x_p(t))\right) . \tag{5}$$

The best aggregation operator must be chosen according to the semantics of the decision. A conjunctive operator, such as the "minimum" or the "product", gives good results for expressing that all criteria must agree. A weighted operator like OWA [9] may be used to express some compromise between partial conclusions. The final decision is computed by a decision rule. The most usual decision rule is the “maximum rule”, where the decision is chosen according to the greatest membership value.

The current approach is flexible enough so that some partial conclusions can be computed from the combination of two or three variables in multi-variable rules. An aggregation operator computes a final conclusion from partial conclusions obtained from all sub-models.

In some applications, such as the one described in this work, it is desirable to get a final score associated to de decision such that intermediate decisions could be analyzed. In those cases a defuzzification step is used to compute the final score, by associating a score vector $\mathbf{w} = [w_1, \ldots, w_m]$ , where the score $w_k$ is associated to the decision $D_k \in \mathbf{D}$ . The final score is computed as:

$$score(t) = \sum_{k=1..m} \mu_{D_k}(\mathbf{x}(t)).w_k \ . \tag{6}$$

where the decision membership value $\mu_{D_k}(\mathbf{x}(t))$ is computed by equation (5).

## 4 The Fuzzy MCDM Approach for Location Selection

The objective of the computational tool developed in this work is to help the decision maker on choosing a landfill site following the three groups of criteria mentioned above: environmental, social and economical. The output of the computational system can furnish partial quality grades for each category, so the characteristics of the sites are monitored following those groups.

The uncertainties in the data are modeled as quantitative and qualitative criteria. The quantitative criteria represent values mainly defined in Brazilian Norms while the qualitative criteria are linguistic values and encode the knowledge of experts. Those criteria involve human subjective judgments that constitute a type of imprecise data that are not easily represented with traditional computing. Thus, fuzzy logic is a more natural approach to deal with those types of problems.

The outputs of the decision-support system mean the qualification degree of the sites for the installation of a new MSW landfill. The decision set $\mathbf{D} = \{D_1 \ldots D_m\}$ represent the linguistic terms {“Bad”, “Acceptable”, “Adequate”}.

The Fuzzy Inference receives the fuzzy, linguistic criteria that are used to activate the rules from the rule base. They are in terms of linguistic variables and have fuzzy sets associated with them. The inference engine thus maps the input fuzzy sets into output fuzzy sets, handling the knowledge in a procedure similar to human reasoning and decision-making. The rule base consists of 60 rules that describe, separately by groups, the involvement of the criteria on the qualification degree that means the final decision. Weights are attributed to each criterion according to the norms. Subsequently, the system furnishes the qualification degree separately by each group and a global rate is provided as a weighted average of the three group rates previously calculated that point toward the best site. The inputs can be weighted to each group based on user’s priorities.

## 5 Case Study

In the following case study, the weights of the three groups were equally distributed and the final score is the average of the scores obtained by each group. Decision makers may find it necessary to furnish different values for these weights, if there is a main concern about a particular group of criteria. This example is a landfill site already operating which had four candidate sites, that were analyzed in this case study. Table 1 shows the input data of the four sites.

**Table 1.** Input data

| Criteria | | Site 1 | Site 2 | Site 3 | Site 4 |
|---|---|---|---|---|---|
| Distance to surface water bodies (m) | | 300 | 1000 | 2000 | 2000 |
| Soil Permeability (cm/s) | | $10^{-6}$ | $10^{-4}$ | $10^{-6}$ | $10^{-6}$ |
| Depth to the ground water level (m) | | 1,0 | 1,0 | 3,0 | 3,0 |
| Distance from airports (km) | | 12 | 15 | 0 | 8 |
| Extension of drainage basin | | Medium | Large | Large | Small |
| Land use | | Not PA | Not PA | Not PA | Not PA |
| Distance from residential areas | | 2000m | 2000m | 100m | 2500m |
| Distance from low-income communities | | Medium | Medium | Small | Large |
| Occupation of access routes | | Medium | Small | Large | Medium |
| Problems with local communities | | Medium | Medium | Medium | Medium |
| Availability of cover material | Distance | Near | Far | Far | Near |
| | Quantity | Small | Small | Small | Large |
| Life time | | 2 years | 15 years | 15 years | 25 years |
| Land cost | | Low | Low | Low | Low |
| Investment in infrastructure | | Small | Medium | Small | Medium |
| Access to heavy trucks | | Medium | Difficult | Easy | Easy |
| Distance from the collect center | | Small | Small | Small | Small |
| Ground slope | | 8% | 3% | 3% | 20% |

The results shown in Table 2 indicate Site 4 as the best place to construct the new landfill. This result is in agreement with the final decision already taken by the design consultants so that the new landfill is already in operation in this site, which validates the fuzzy decision making system. Site 3 presents the worst performance in the social group. Definitely, it is very close to residential areas and to low-income communities. Site 1 presents the poorest performance in the environmental group, mainly due to its proximity to surface water bodies. In this case, the final decision for the best site is very much straightforward as far as Site 4 shows not only the best final grade but also the best partial grades for all three groups. Consequently, the decision maker has a clear view without any extra effort toward a more detailed analysis of the partial grades. In some situations, it does not turn out this way, which obliges decision makers to continue on a more detailed analysis of output results.

**Table 2.** Results

| Group | Partial Grades | | | |
|---|---|---|---|---|
| | Site 1 | Site 2 | Site 3 | Site 4 |
| Environmental | 1,731 | 2,253 | 5,000 | 6,154 |
| Social | 5,000 | 6,000 | 1,167 | 6,722 |
| Economics | 4,365 | 5,802 | 7,654 | 8,889 |
| **Final grade** | **3,699** | **4,685** | **4,607** | **7,255** |

## 6 Final Remarks

The computational system developed in this work is a friendly tool, designed to help the decision maker in choosing a landfill site among potential available areas based on some criteria. It employed a Fuzzy Inference (FI) engine that mapped the input fuzzy sets into output fuzzy sets, handling the knowledge in a procedure similar to human reasoning and decision-making. The rule base described the involvement of the input criteria on the qualification degree of the sites, meaning the output of the decision-support system. The case studied was a landfill site already operating and it was able to validate the system showing that it worked very well and the results agreed with the final decision taken previously by the design consultants. Finally, the developed system has shown to be an efficient tool that not merely helps decision-makers in choosing a landfill site but can also give an overall picture of the available sites.

## References

1. Moula, M., Toll, D.G., Vaptismas, N.: Knowledge-based systems in geotechnical engineering. Geotecnique 45(2), 209–221 (1995)
2. Dorner, S., Shi, J., Swayne, D.: Multi-objective modelling and decision support using a Bayesian network approximation to a non-point source pollution model. Environmental Modelling & Software 22(2), 211–222 (2007)
3. Hepting, D.H.: Decision support for local environmental impact assessment. Environmental Modelling & Software 22(3), 436–441 (2007)
4. Sadiq, R., Husain, T.: A fuzzy-based methodology for an aggregative environmental risk assessment: a case study of drilling waste. Environmental Modelling & Software 20(1), 33–46 (2005)
5. Al-Jarrah, O., Abu-Qdais, H.: Municipal solid waste landfill sitting using intelligent system. Waste Management 6(5), 534–546 (2006)
6. Cho, K.T.: Multicriteria decision methods: an attempt to evaluate and unify. Mathematical and Computer Modelling 37, 1099–1119 (2003)
7. Dubois, D., Prade, H., Testemale, C.: Weighted fuzzy pattern matching. Fuzzy Sets and Systems 28, 313–331 (1988)
8. Evsukoff, A.G., Ebecken, N.F.F.: Mining fuzzy rules for a traffic information system. In: Palade, V., Howlett, R.J., Jain, L. (eds.) KES 2003. LNCS, vol. 2773, pp. 237–243. Springer, Heidelberg (2003)
9. Yager, R.R.: On ordered weighted averaging aggregation operators in multi-criteria decision making. IEEE Trans. Systems Man Cybernetic 18, 183–190 (1988)

# Fuzzy-Based Call Admission Control Scheme for Mobile Networks

Jing-Long Wang and Shu-Yin Chiang

Department of Information and Telecommunications Engineering,
Ming Chuan University, Taipei 11120, Taiwan
jlwang@mcu.edu.tw

**Abstract.** The fuzzy-based call admission control (CAC) scheme is presented in this paper to offer adaptive services for multimedia stream in next generation wireless networks. Since the contemporary wireless networks will provide a variety of services and each service has multiple levels of quality requirements, but the resource is always scarce. Thus, the adaptive resource allocation for bandwidth is an important issue. In this paper, the proposed CAC scheme considers the adaptive resource requirements to enhance the efficiency of channel utilization in wireless networks. When a new call is arriving, the CAC scheme will evaluate if the available bandwidth can satisfy the requirement of incoming call. Whenever the available bandwidth is not sufficient to meet the requirement, the CAC scheme is based on fuzzy logic to choose an existing call, of which the allocated bandwidth will be degraded in order to release some bandwidth for the incoming call. From the results of simulation, the proposed scheme is superior to the existing scheme.

**Keywords:** Fuzzy, CAC, QoS, Wireless Networks.

## 1 Introduction

Recently, as the increased demand of bandwidth capacity and QoS for multimedia, the main resource constraint in the wireless network is the bandwidth available for transmission, due to the inherent bandwidth scarceness. Moreover, the traffic load changes drastically with time and position and thus the demand of channels is dependent of time and position of cells. Therefore, it is important to develop an effective method for efficiently assigning bandwidth to each call [1].

There are two main parameters, the blocking rate and the dropping rate, that are used to evaluate the quality of service in wireless networks [2]. The blocking rate denotes the rejection rate of new calls, while the dropping rate denotes the cancellation rate of handoff calls. The new call is the new initiated call, while the handoff call is the existing and working call that is going to transfer from one cell to the other cell due to the movement of mobile station. Obviously, to interrupt a working call, the handoff call, will bring more inconvenience than to disallow a new initiated call, the new call. Thus, most channel assignment methods treat the handoff call with higher priority than the new call. In addition to protect handoff call, an effective method has

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 958–965, 2008.

to offer the required QoS for mobile stations as well as refine the utilization of bandwidth. In order to provide the adaptive QoS provision, each connection that requires QoS guarantee is given an adaptable profile according to its traffic characteristics when the connection is initialized, including the adaptable range of QoS requirements such as maximum and minimum QoS requirements. In the stage of connection establishment, the BS (base station) employs the call admission control to determine whether to accept or reject a new connection according to available bandwidth and the adaptable profile. During the run-time period, the BS has to monitor both changes of each connection and the whole network system, and dynamically regulate the bandwidth allocation of wireless spectrum according to QoS constraints of each call [3, 4].

In next generation wireless networks, CDMA has emerged as one of the most promising multiple accesses and widely adopted fourth-generation air interface. In this paper, a new adaptive CAC method for CDMA system are presented to satisfy differentiated QoS requirements as well as to maximize high system utilization. The proposed method firstly determines if there is available bandwidth for an incoming call. If there is not sufficient bandwidth for the incoming call, it employs the fuzzy logic to decide which on-going call is able to reduce the quantity of occupied bandwidth. In this way, more calls can be accepted to work in wireless networks, while the existing calls only reduce some tolerable bandwidth. From the simulation results, the proposed method is able to carry more connections and improve system utilization.

There are five sections in this paper. In the section 2, the CAC methods of previous works are introduced. Some comparisons of their advantages and disadvantages are also illustrated. The section 3 presents the proposed method. The section 4 describes the system model and the results of the simulation. Finally, a conclusion is given in the section 5.

## 2 Previous Works

CAC methods have a great impact on the efficiency and performance of system throughput. The design of an efficient resource management for CAC method is a difficult task that typically involves many conflicting considerations which have to be analyzed to find a smooth and balanced solution. A lot of CAC methods have been proposed to satisfy various QoS requirements.

A multimedia QoS provision exploits the concepts of Static Priority (SP) and Minimum Set (mSet). Each component within a multimedia is assigned a significance at call set-up stage that indicates a level of SP. For example, a videophone application contains two media, including voice and video, in which reasonably voice component has higher SP than video components. Minimum Set indicates the minimum bound of QoS requirements. For example, if a teleconference session contains voice, audio, and video, users may decide not to progress the call when audio or video cannot be transmitted.

In [3], it mainly proposes an algorithm to allocate the resources at CAC period. Each connection is assigned an acceptable range of transmission rate and also preprioritized according to its traffic characteristics.

An adaptive scheme of penalty-based adaptable reservation and admission (PARA) was presented in [4]. Each connection may contain more than one set of QoS requirement, in which one for normal mode and others for degrade modes. With the penalty information, the BS based on the traffic conditions provides acceptable service quality to all the connections by minimizing the aggregate penalty. The BS accepts a new connection only if the penalty of blocking the new connection is larger than the sum of penalty caused by the degradation of other existing connections.

## 3 Adaptive QoS Scheme

Basically, mobile stations talking to each other are going through a base station. A mobile device can send a QoS request to the base station while a new call is issued. In the QoS request, there could have a lot of valuable parameters, including the required bandwidth and the priority level. When the CAC mechanism receives this request from a mobile device, these parameters will be evaluated in order to determine whether the incoming call is accepted or not. Whenever the available bandwidth is currently not available to satisfy the requirement of incoming call, the degradation scheme is employed to dynamically adjust the bandwidth allocated to the existing calls. The occupied bandwidth of an existing call can be reduced such that some of bandwidth can be released for reallocating to the new incoming call.

The priority of a call is denoted as $P_i$, in which the higher the value of i, the higher the priority. In the QoS definition of 3GPP, there are four classes of priority, including P1, P2, P3, and P4, representing the background, the interactive, the streaming, and the conversation service class, respectively. Among these four classes, the conversion service class has the highest priority, while the background service class has the lowest priority.

In addition, the bandwidth required by a call may be different, which is indicated by $B_i$. Similarly, the higher value of i represents the higher bandwidth requirement. In this paper, it is assumed that every call at least should have two kinds of bandwidth requirement, in which $B_{min}$ indicates the minimum bandwidth requirement, and $B_{max}$ indicates the maximum bandwidth requirement.

## 4 Fuzzy Logic Degradation CAC Algorithm

The proposed scheme perform the bandwidth degradation decision is based on the fuzzy logic and three important factors, including the resource, the bandwidth and the priority. In this section, the fuzzy logic degradation scheme is described, in which there are three main steps, including the fuzzification, the rule evaluation, and the defuzzification.

### 4.1 Fuzzification

The first step of the proposed degradation scheme is fuzzification, in which the membership function should be defined. Firstly, two input variables and four membership functions are defined. Two input variables are BW(Bandwidth) and PC(Priority Class). Four membership functions are MBW(Maximum BW), NBW(Minimum BW), HPC(High PC), and LPC(Low PC). The membership values can be obtained by applying the values of input variables to different membership functions. Assume the membership values under consideration are generalized from triangular membership functions, in which the membership value is located at the range from 0 to 1. The triangular membership function $\mu_A(x)$ is defined by

$$\mu_A(x) = \begin{cases} \frac{(x-a_1)h_a}{a_2-a_1}, & a_1 \le x \le a_2 \\ \frac{(a_3-x)h_a}{a_3-a_2} & a_2 \le x \le a_3 \\ 0 & otherwise \end{cases}, \tag{1}$$

where $0 \le h_a \le 1$.

## 4.2 Rule Evaluation

The fuzzy rules are defined according to the membership value, in which there are a lot of IF-THEN rules for deciding which path should be chosen. Three membership values can be used to produce eight combinatorial rules. The fuzzy number (FN) of each rule is defined as the minimum value of the two membership values, as shown in equation (2).

$$\text{FNi}=\min\{\text{MBW/NBW, HPC/LPC}\}. \tag{2}$$

FN is the minimum value of each combination of membership values, while i is the value from 1 to 4. If MBW=0.2, and HPC=0.5, the fuzzy number will be 0.2. Each rule may have four kinds of different results, including Yes(Y), Probably Yes (PY), Probably No (PN), and No (N). Thus, the bandwidth degradation decision (BDD) can be defined as the largest fuzzy number of individual results, as shown in the equation (3), in which the value of p is one of the four possibilities, including Y, PY and PN or N.

$$\text{BDD(p)}=\text{Max}\{\text{FN1, FN2,.......,FNt}\}. \tag{3}$$

## 4.3 Fuzzy Logic Routing Decision

In the phase of defuzzification, the process of transforming the result of fuzzy inference into an exact number is called fuzzy decision (FD). A weight value is set for each possible result, including Y, PY, PN and N. The final fuzzy decision can be obtained by the equation (4), in which the value of FD is between 0 and 1. The call with the minimum value of FD will be the most feasible choice.

$$FD = \frac{\sum M_m \times W_m}{\sum M_m} \tag{4}$$

$M_m$: $BDD_m$ value,

$W_m$: the weight value for the result m,

m: Y, PY, PN or N.

## 4.4 Fuzzy Logic Degradation Model

Before the proposed model is introduced, some parameters are defined. $B_{avail}$ is the total available bandwidth. $B_{max}$ is the required maximum of bandwidth of a call in the cell and $B_{min}$ is the acceptable minimum bandwidth of a call. $D_{fd}$ is fuzzy decision that decides whether the allocated bandwidth will be degraded or not. $T_{rel}$ is threshold value and $D_{fd}$ is less than $T_{rel}$ that the call can be degraded. When a new call or a handoff call arrives, the CAC is triggered by the request of mobile terminal. If the available bandwidth is abundant and greater than the maximum requirement of an arriving call, the maximum bandwidth would be allocated to this arriving call. However, when the cell is fully loaded, one or some of the existing calls may be degraded to minimum. The fuzzy decision algorithm will choose the candidate calls that are able to release some bandwidth for the arriving call. The CAC and fuzzy algorithms are shown in the following

```
CAC()  /* Call Admission Control */
{
while ( incoming($call_i$) )
{
        if ( $B_{avail}$ > $B_{max}$($call_i$) )
         Allocate($call_i$, $B_{max}$);
        else if ( $B_{avail}$ > $B_{min}$($call_i$) )
         Allocate($call_i$, $B_{min}$);
        else
         {
          $B_{avail}$ =Fuzzy($B_{min}$);
          if ( $B_{avail}$ > $B_{min}$($call_i$) )
           Allocate($call_i$, $B_{min}$);
          else
           Block( $call_i$ );
         }
       }
}

fuzzy($B_{min}$)
{
       while ($B_{avail}$ < $B_{min}$)
       {
        k = min{ $FD_k = \frac{\sum M_k \times W_k}{\sum M_k}$ };

        if (k = 0)
return(0);
{
$B_{avail}$ = $B_{avail}$ +(B(k) - $B_{min}$(k));
B(k) = B(k)- $B_{min}$(k);
        }
       }
       Return($B_{avail}$)
}
```

## 5 Simulations

The wireless network with 64 cells is studied to investigate the performance, as shown in Figure 1. Each base station is assumed to have 32 channels. The arrival rates

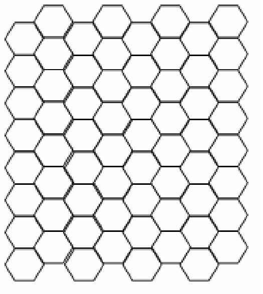


**Fig. 1.** The simulated wireless network with 64 cells

of new calls ( $\lambda_n$ ) are assumed to be 0.1~0.8 (call/sec). The arrival rates of handoff calls ($\lambda_h$ ) are in the range of 0.08~0.25 (call/sec). The service rates of new calls ($\mu_n$ ) and handoff calls ($\mu_h$ ) are 1/60 (call/sec) and 1/60 (call/sec). The mean serve time of a call is 60 seconds. It is assumed that the tolerable dropping rate (*ptd*)is $10^{3}$.

The fuzzy method is compared with the fixed allocation methods, including the $B_{max}$ and $B_{min}$ allocation methods. The $B_{max}$ method is based on the maximum bandwidth to assign the required bandwidth, while the $B_{min}$ method is based on the minimum bandwidth to assign the required bandwidth. Figure 2 show the bandwidth utilization for different arrival rates of handoff calls when the arrival rate of new call is from 0.1 to 0.8. From the results, the $B_{min}$ method has the worst bandwidth utilization, while the $B_{max}$ method and the fuzzy method have the better performance for the utilization. However, when the arrival rate is higher than 0.6, the utilization of the $B_{max}$ method cannot have any further improvement, while the fuzzy method is still able to accept new calls and so the utilization can be increased continuously.

Figure 3 illustrates the dropping rate versus arrival rate. Evidently, the fuzzy method and $B_{min}$ method are capable of controlling the dropping rate under the tolerable dropping rate ($10^{-3}$), while $B_{max}$ method cannot satisfy the requirement of the tolerable dropping rate when the arrival rates of handoff calls become larger. This is due to the fact that the fuzzy method can adaptively degrade the bandwidth of some existing calls, and thus result in the lower dropping rate.

Figure 4 presents the blocking rate versus the arrival rate of handoff calls. The blocking rate increases with the arrival rate. This is due to that most bandwidth will be occupied by the handoff calls. As a result, new calls cannot acquire the required

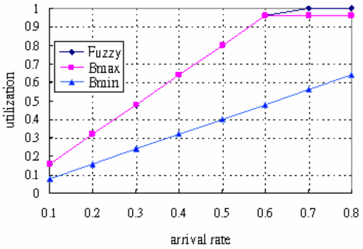


**Fig. 2.** Utilization versus Arrival Rate

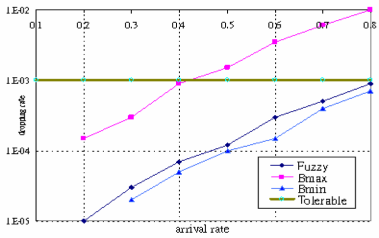


**Fig. 3.** Dropping Rate versus Arrival Rate

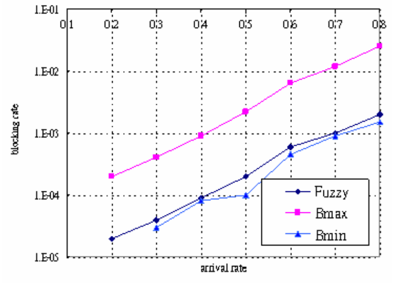


**Fig. 4.** Blocking Rate versus Arrival Rate

bandwidth. From the results, the blocking rate of the fuzzy method and the $B_{min}$ method are better than that of the $B_{max}$ method. This is because the $B_{max}$ method allocates too much bandwidth for existing calls, and therefore less bandwidth can be allocated to new calls. On the other hand, the fuzzy method adaptively allocates the bandwidth. When the arrival rate is high and the available bandwidth is small, the fuzzy method will release some bandwidth that originally are allocated to existing calls, and then reallocate the released bandwidth to new calls. Thus, the blocking rate can be effectively reduced.

## 6 Conclusions

In this paper, a new CAC method is proposed for mobile networks. This method based on the fuzzy logic takes into account both the bandwidth requirement and the priority level. The fuzzy method adaptively allocates the bandwidth. When the arrival rate is high and the available bandwidth is small, the fuzzy method will release some bandwidth that originally are allocated to existing calls, and reallocate the released bandwidth to new calls. In addition, the proposed method provides QoS guarantee by keeping the handoff dropping rate below the desired level. The simulation results show that the proposed scheme can effectively overcome the problems that the traffic load changes very fast and is distributed non-uniform in the wireless network. Moreover, the proposed method is capable of confining the dropping rate below the predefined limitation, and enhancing the blocking rate significantly in comparison with other methods.

## References

1. Wang, J.-L., Chiang, S.-Y.: Adaptive Channel Assignment Scheme for Wireless Networks. Journal of Computers and Electrical Engineering 30(6), 417–426 (2004)
2. Chou, C.-T., Shin, K.G.: Analysis of Adaptive Bandwidth Allocation in Wireless Networks with Multilevel Degradable Quality of Service. IEEE Transactions on Mobile Computing 3(1), 5–17 (2004)
3. Wang, J.-L., Chen, C.-H.: Adaptive Two-stage QoS Provisioning Schemes for CDMA Networks. Computer Systems Science and Engineering 22(6), 56–64 (2007)
4. Wang, J.-L., Chen, C.-W.: Fuzzy Logic Based Mobility Management for 4G Heterogeneous Networks. In: Gabrys, B., Howlett, R.J., Jain, L.C. (eds.) KES 2006. LNCS (LNAI), vol. 4252, pp. 888–895. Springer, Heidelberg (2006)

# Routing Analysis Using Fuzzy Logic Systems in Wireless Sensor Networks

Shu-Yin Chiang and Jing-Long Wang

Department of Information and Telecommunications Engineering,
Ming Chuan University
5 Deh-Ming Road, Gwei-Shan, Taoyuan 333, Taiwan
sychiang@mcu.edu.tw

**Abstract.** This paper presents a novel algorithm for routing analysis in wireless sensor networks utilizing a fuzzy logic system at each node to determine its capability to transfer data based on its relative energy levels, distance and traffic load to maximize the lifetime of the sensor networks. The fuzzy logic system helps for the selection of node to forward packets to the destination. The simulation results show that network lifetime can be improved by employing the proposed routing algorithm.

**Keywords:** wireless sensor networks (WSN), fuzzy logic systems (FLS).

## 1 Introduction

Wireless sensor network (WSN) is composed of cheap and tiny unreliable sensors with limited resources, where the sensors possess sensing, computing and communicating capabilities. Due to recent advances in the integrated circuit and micro-electro-mechanical systems and wireless communications, WSN is expecting a rapidly increasing uptake in various applications. However, sensor nodes are often battery-operated, once deployed they work until the energy is depleted. Given the high density of typical WSNs and their deployment in remote or hostile areas, manual replacement of nodes is unfeasible. Therefore, networks are normally deployed for data collection where human intervention after deployment, to recharge or replace node batteries may not be feasible, resulting in limited network lifetime. The network lifetime is dictated by the duration of individual nodes, making the energy saving a crucial requirement. Moreover, the data collection according to the network structure and remaining values of the node energy is different from the traditional computer networks. The design constituent of the routing protocol depends mainly on the application because of the application's traffic demand and pattern may vary enormously. Power consumption, mobility, scalability and QoS [2-8] are the other most significant issues in designing routing protocols in WSN. To extend the sensor network lifetime, we utilize the Fuzzy Logic System (FLS) that optimizes the routing path in a distributed fashion.

Our algorithm explores fuzzy logic as a solution. Fuzzy Logic Systems (FLS) are in general non-linear input–output mappings [1]. FLS operates with fuzzy sets, which

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 966–973, 2008.

extends the ordinary notion of crisp sets. A fuzzy set $F$ is characterized by a membership function, which gives the degree of similarity of $x$ to $F$. In engineering, the most widely used are the rule-based FLS. These systems are composed of four basic components as shown in Figure 1. First, the fuzzifier maps crisp inputs into fuzzy sets by using the membership functions. Second, the fuzzified values activate the rules, which provided by the experts or extracted from numerical data. The rules are expressed as a collection of IF-THEN statements, having fuzzy propositions as antecedents and consequences. Third, the fuzzy inference engine combines the rules to obtain an aggregated fuzzy output. Finally, the defuzzifier maps the fuzzy output back to a crisp number that can be used for making decisions or control actions.

In this paper, we propose a novel routing analysis algorithm to automatically select the data dissemination protocol that better meets application-specific requirements while minimizing the network resource consumption. The remaining paper is organized as follows. Section 2, describes some related works. Section 3 states the fuzzy logic algorithm. In Section 4, we present our simulation results and finally in Section 5, the conclusions are stated.

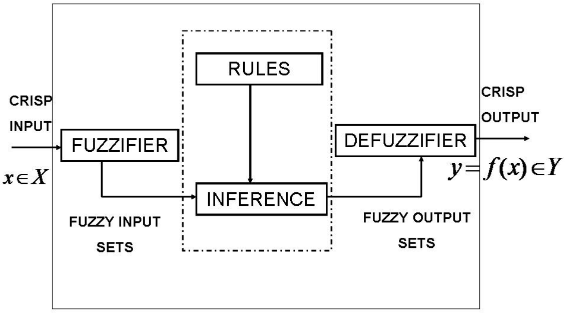


**Fig. 1.** The structure of a fuzzy logic system

## 2 Related Work

Routing in sensor networks involves finding the optimal transmission path for the energy constrained sensor nodes to the destination in order to prolong the network lifetime. From the aforementioned literatures [2-3], we find some criterion to lengthen the lifetime of the sensor networks as follows:

- Small multiple hops: As the energy consumed for the transmission is proportional to the square of the distance from sender to receiver, multiple short hops is preferable instead of a single large hop.
- Shortest path: Shortest path from the sender to receiver is the straight line connecting the nodes. Forwarding packets along this line is more efficient than a detour.
- Traffic load: In case, concentration of events in some particular areas is more than that of other areas, using shortest path will cause implosion along the path. Therefore, the traffic load in the nodes will effect the lifetime of the networks.

- Energy: Nodes having greater remaining energy participates more than the nodes having small amount of power can extent the network lifetime.

Energy consumption is the key issue in wireless sensor network. Fuzzy diffusion is introduced in [4] to obtain an energy optimization on the directed diffusion scheme by proposing an energy optimization that can be incorporated into existing diffusion schemes: Fuzzy Diffusion shifts the energy cost of data forwarding to non-critical nodes (nodes having high residual energy or less data), while still achieving an energy-balance in the network. Critical nodes (low energy nodes with heavy traffic) have reduced data forwarding burden and expend most of their power in sensing and communicating their sensor data, thus seeking to achieve net longevity. Fuzzy diffusion will be ideal for surveillance applications, where sensor nodes are densely deployed for sustained observation of physical events. The two linguistic input variables used by the fuzzy controller are:

***Relative Energy Level (REL)*** of a node, defined as the residual energy status of a node with respect to its neighborhood. This factor represents a node's energy criticality and is given by:

$$REL = \frac{E_{node} - E_{\min}}{E_{\max} - E_{\min}}, \tag{1}$$

where $E_{max}$ and $E_{min}$ are maximum and minimum energy levels in the neighborhood and $E_{node}$ is the node's residual energy level. The *REL* definition is an approximate rank function that indicates the energy ranking of a node among its neighbors. Higher the *REL*, lesser is the energy-criticality of a node. Nodes exchange energy information by piggybacking their current energy levels along with interest and data messages and each node maintains a cache for storing neighbor residual energy levels.

***Traffic Intensity (TI)*** of a node, defined as the amount of traffic pending in a node's queue. This includes the application traffic and also the traffic that a node has already committed to forwarding. This factor represents the *traffic burden* on a node and is given by the ratio between the traffic in the node's queue and the maximum queue size of the node. .Lower the *TI*, lesser the load in a node. Both *REL* and *TI* lie in the range 0 to 1 and jointly define the criticality of a node.

Combining the above reference papers, we will introduce a Fuzzy Logic System by using the distance, energy and traffic load to deal with the routing analysis in wireless sensor networks.

## 3 Fuzzy Logic Algorithm

The Fuzzy Logic Algorithm is illuminated by the powerful capability of fuzzy logic system to handle uncertainty and ambiguity. Fuzzy logic system is well known as model free. Their membership functions are not based on statistical distributions. In this paper, we apply fuzzy logic system to optimize the routing process by some criterion. The main goal is designing the algorithm to use Fuzzy Logic Systems to lengthen the lifetime of the sensor networks.

### 3.1 Input Variables

There are four following fuzzy input variables used by the Fuzzy Logic

- Energy Level (*P*)

The energy remaining level is defined as follows:

$$P = \frac{E_{node} - E_{\min}}{E_{\max} - E_{\min}}, \tag{2}$$

where the $E_{max}$ and $E_{min}$ are the maximum and minimum energy level in the neighborhood, respectively, and $E_{\text{node}}$ is node's remaining energy level. Higher the *P* value, lesser the energy-criticality of a node. Nodes exchange energy information by piggy backing their current energy levels along with interest and data messages, and each node maintains a cache for storing neighbor residual energy levels.

- The distance from Node (*d*)

The transmitting power of a node is proportional to the square of the distance between the candidate nodes to the source node. Therefore, the distance from source node location ($n_{source}$) to the location of the destination node ($n_{destination}$) is defined as follows:

$$d = \left| n_{source} - n_{candidate} \right| \tag{3}$$

- Distance from the shortest path ($d_{min}$)

Not only the distances of nodes will be calculated but also needs to find the shortest path. The shortest path is defined in the following:

$$d_{\min} = \min \left| n_{source} - n_{candidate} \right| \tag{4}$$

- Traffic Load (*TL*)

The traffic load of a node, defined as the amount of traffic pending in a node's queue. This includes the application traffic and also the traffic that a node has already committed to forwarding. The factor represents the traffic load on a node and is given by:

$$TL = \frac{Traffic\ in\ node's\ queue}{Maximum\ queue\ size\ of\ the\ node} \tag{5}$$

Lower the TL value, lesser the load in the queue.

### 3.2 Membership Functions

The first step of designing fuzzy algorithm requires characterizing the membership function (MF), which gives the input output relations. Membership functions are different for the different metrics. The input parameters are the routing metrics (x-axis) with respect to the corresponding cost (y-axis) of the MF and the outputs are projected to form the trapezoids. From the fuzzy input variables as mentioned above, we define the member functions as follows:

- Energy Level (*P*)

The energy remaining level is defined in equation (2). For the Energy remaining level the MF is set in *F(1)=1-P*. The lower energy remaining level shows the highest resistance to forwarding a packet. The *F(1)* indicates the first membership function.

- Distance from the node (*d*):

As mentioned above that the power is proportional to the square of the distance, in case of the first order radio model [8] (see Figure 2), the distances are the inputs of the MF. Outputs, the projected trapezoids are the weight for the corresponding node. The height of the trapezoid for the node is defined as square of the *d*, *F(2)=normalized* $d^2$. The *F(2)* indicates the second membership function.

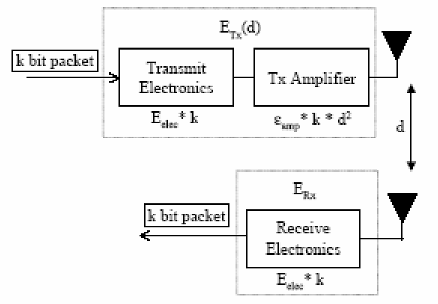


**Fig. 2.** First Order Radio Mode

- Distance from the shortest path ($d_{min}$)

The MF, in this case, is the same as the previous one because it is also a distance. Inputs are the $d_{min}$ and the outputs are the corresponding trapezoids. The height of the trapezoid for the node is defined as square of the $d_{min}$, *F(3)= normalized* $d_{min}^2$. The *F(3)* indicates the third membership function.

- Traffic Load (*TL*)

The Traffic load is defined in equation (5). We use a linear function as a traffic load of a node in MF. The heavier of the traffic load then the more it becomes reluctant to forward the packet. The height of the trapezoid for the node is defined as *TL*, *F(4)=TL*. The *F(4)* indicates the fourth membership function.

**Decision:** From above all the four types of outputs are added and the weighted average is taken. The area of the trapezoids, are calculated by the following expressions.

$$A = \sum_{i=1}^{4} w_i \frac{(1-(1-F(i)^2)}{2}, \tag{6}$$

where *A* denotes weighted averages the area of the trapezoid and *F* is the membership function and the $w_i$ is the weighting parameter. The higher value *A* indicates that it is the node that source will forward to due to its minimum cost.

Consequently we take four routing parameters to make decisions in the membership function. These parameters can be easily measured through localization or GPS (e.g. distance), power level access etc and are required to be periodically updated through single hop broadcasts.

## 4 Performance Evaluation

To evaluate the performance of the protocol, we simulate the protocol in MATLAB. In our simulation we deploy 100 sensor nodes in a 100mx100m field as shown in Figure 3. We use the radio model introduced in the area of routing protocol evaluation in WSN [2, 3, 8]. In this model, the transmission and receive costs are characterized through the data rate and distances. In our work, we assume a simple first order radio model where the radio dissipates $E_{elec}$= 50 nJ/bit to run the transmitter or receiver circuitry and $\varepsilon_{amp}$ = 100 pJ/bit/m$^2$ for the transmit amplifier (see Figure 2). We make the assumption that the radio channel is symmetric such that the energy required to transmit a message from node A to node B is the same as the energy required to transmit a message from node B to node A for a given SNR.

We simulate the system where the traffic is generated randomly with equal distribution. To observe the performance of the proposed scheme, we consider the number of transmissions versus the number of dead nodes by using the fuzzy logic algorithm and the conventional routing algorithm in the wireless sensor network system.

From the Figure 4, we find out that the proposed scheme (on solid line) performs much better than the conventional system only energy considered (dashed line). The

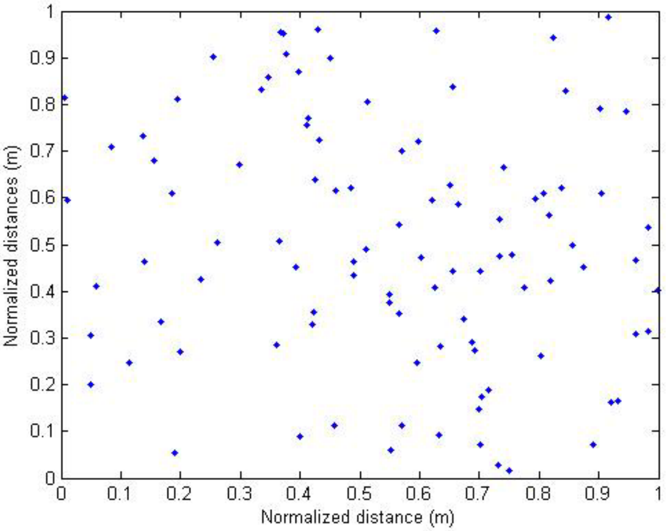


**Fig. 3.** Sensor deployment

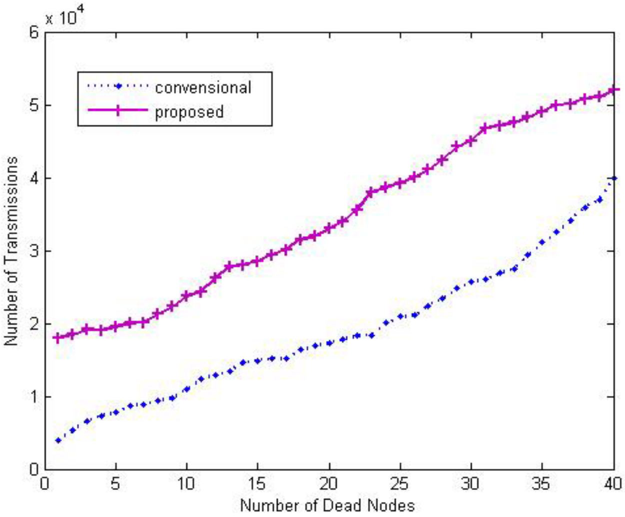


**Fig. 4.** Simulation results by using the FLS versus the conventional system

reason is that the fuzzy logic system causes the fair distribution of the traffic load and the energy among the nodes.

## 5 Conclusions

This paper describes a fuzzy logic system for routing analysis to prolong the sensor node lifetime in wireless sensor networks. A routing decision is made by each node based on the output of fuzzy logic system. These include the distance, traffic load and energy of nodes. Simulation results show that the networks lifetime could be extended by the proposed scheme.

## References

1. Mendel, J.M.: Fuzzy Logic Systems for Engineering: a Tutorial. IEEE Prcoeeedings 83(3), 345–377 (1995)
2. Azim, M.A., Jamalipour, A.: Optimized Forwarding for Wireless Sensor Networks by Fuzzy Inference System. In: IEEE International Symposium on Wireless Broadband and Ultra Wideband Communications, Sydney (2006)
3. Azim, M.A., Jamalipour, A.: Performance Evaluation of Optimized Forwarding Strategy for Flat Sensor Network. In: IEEE International Symposium on Global Telecommunications, pp. 710–714 (2007)
4. Balakrishnan, M., Johnson, E.E.: Fuzzy Diffusion for Distributed Sensor Networks. In: IEEE International Symposium on Military Communications, Atlanta City, NJ, vol. 5, pp. 1–6 (2005)

5. Munir, S.A., Bin, Y.W., Biao, R., Jian, M.: Fuzzy Logic based Congestion Estimation for QoS in Wireless Sensor Network. In: IEEE International Symposium on Wireless Communications and Networking, pp. 4336–4341 (2007)
6. Lazzerini, B., Marcelloni, F., Vecchio, M., Croce, S., Monaldi, E.: A Fuzzy Approach to Data Aggregation to Reduce Power Consumption in Wireless Sensor Networks. In: Fuzzy Information Processing Society, pp. 436–441 (2006)
7. Pirmez, L., Delicato, F.C., Pires, P.F., Mostardinha, A.L., de Rezende, N.S.: Applying fuzzy logic for decision-making on Wireless Sensor Networks. In: IEEE International Symposium on Fuzzy Systems Conference, pp. 1–6 (2007)
8. Heinzelman, W., Chandrakasan, A., Balakrishanan, H.: Energy-Efficient Routing Protocols for wireless microsensor networks. In: 33rd Hawaii Int. Conference System Sciences (HICSS), pp. 3005–3014 (2000)

# Prioritization of Incomplete Fuzzy Preference Relation

Hsuan-Shih Lee[1], Pei-Di Shen[2], and Wen-Li Chyr[3]

[1] Department of Shipping and Transportation Management
National Taiwan Ocean University
Keelung 202, Taiwan
[2] Graduate School of Education
[3] Department of Information Management
Ming Chung University
Taipei 111, Taiwan

**Abstract.** The concept of incomplete fuzzy preference relation is discussed in this paper. We focus on additive consistent incomplete fuzzy preference relations, in which the correspondence between priority vector and incomplete fuzzy preference relation is investigated. We find that an additive consistent incomplete fuzzy preference relation does not imply there is a priority vector that satisfies additive transitivity.

**Keywords:** consistent incomplete fuzzy preference relation, incomplete fuzzy preference relation, priority vector.

## 1 Introduction

Introducing of the fuzzy theory into decision model is one of advancements that facilitate the decision process for decision makers. Different models have been proposed for decision-making problems under fuzzy environment [15-18]. Decision-making process usually consists of multiple individuals interacting to reach a decision. Different experts may express their evaluations by means of different preference representation formats and as a result different approaches to integrating different preference representation formats have been proposed [1,2,8,10,29,30]. In these research papers, many reasons are provided for fuzzy preference relations to be chosen as the base element of that integration.

One important issue of fuzzy preference relation is that of “consistency” [3,4,11]. Many properties have been suggested to model transitivity of fuzzy preference relations and some of these suggested properties are as follows:

(1) Triangle condition [11,19]
(2) Weak transitivity [11,23]
(3) Max-min transitivity [11,28]
(4) Max-max transitivity [7,11,28]
(5) Restricted max-min transitivity [11,23]

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 974–979, 2008.

(6) Restricted max-max transitivity [11,23]
(7) Additive transitivity [11,19,23]
(8) Multiplicative transitivity [11,22,23,25,27]

Amongst these properties two of them attract more attentions in recent research [11,25,26], which are additive transitivity and multiplicative transitivity.

One of research focuses of fuzzy preference relations is to prioritize alternatives based on the fuzzy preference relation. Many methods have been proposed to draw priorities from a multiplicative preference relation, such as the eigenvector method [22], the least square method [12], gradient eigenvector method [5], logarithmic least square method [6] and generalized chi square method [24], etc. When using fuzzy preference relations, some priority methods have been given using what have been called choice functions or degrees [1,9,13,14,20,21].

Another important research issue of fuzzy preference relations is how to draw consistent preferences when the fuzzy preference relations are incomplete. Xu [26] has proposed two goal programming models, based on additive consistent incomplete fuzzy preference relation and multiplicative consistent incomplete fuzzy preference relation, for obtaining the priority vector of incomplete fuzzy preference relations. In Xu's method based on additive consistency, he postulated a correspondence between priority vector and additive consistent incomplete fuzzy preference relation. We are going to show that the correspondence is incorrect.

## 2 Preliminaries

For simplicity, we let $N = \{1,2,\ldots,n\}$.

**Definition 2.1.** Let $R = (r_{ij})_{n\times n}$ be a preference relation, then $R$ is called a fuzzy preference relation [2,13,23], if

$$r_{ij} \in [0,1], \qquad r_{ij} + r_{ji} = 1, \qquad r_{ii} = 0.5 \qquad \text{for all } i, j \in N .$$

**Definition 2.2.** Let $R = (r_{ij})_{n\times n}$ be a fuzzy preference relation, then $R$ is called an additive consistent fuzzy preference relation, if the following additive transitivity (given by Tanino [23]) is satisfied:

$$r_{ij} = r_{ik} - r_{jk} + 0.5, \qquad \text{for all } i, j, k \in N$$

Xu extended the concepts in previous section to the situations where the preference information given by the DM (decision maker) is incomplete.

**Definition 2.3. [26]** Let $R = (r_{ij})_{n\times n}$ be a preference relation, then $R$ is called an incomplete fuzzy preference relation, if some of its elements cannot be given by the DM, which we denote by the unknown number $x$, and the others can be provided by the DM, which satisfy

$$r_{ij} \in [0,1], \qquad r_{ij} + r_{ji} = 1, \qquad r_{ii} = 0.5 .$$

**Definition 2.4. [26]** Let $R = (r_{ij})_{n\times n}$ be an incomplete fuzzy preference relation, then $R$ is called an additive consistent incomplete fuzzy preference relation, if all the known elements of $R$ satisfy the additive transitivity

$$r_{ij} = r_{ik} - r_{jk} + 0.5 .$$

For the convenience of computation, Xu constructed an indication matrix $\Delta = (\delta_{ij})_{n\times n}$ of the incomplete fuzzy preference relation $R = (r_{ij})_{n\times n}$, where

$$\delta_{ij} = \begin{cases} 0, & r_{ij} = x \\ 1, & r_{ij} \neq x. \end{cases}$$

Xu [26] developed a goal programming model based on additive consistent incomplete fuzzy preference relation for obtaining the priority vector of incomplete fuzzy preference relation.

Let $w = (w_1, w_2, \ldots, w_n)^T$ be the priority vector of the incomplete fuzzy preference relation $R = (r_{ij})_{n\times n}$, where $w_i \geq 0$, $i \in N$, $\sum_{i=1}^{n} w_i = 1$.

Xu postulated that

(1) If $R = (r_{ij})_{n\times n}$ is an additive consistent incomplete fuzzy preference relation, then such a preference relation is given by

$$\delta_{ij} r_{ij} = \delta_{ij}[0.5(w_i - w_j + 1)], \qquad i, j \in N . \tag{1}$$

Based on (1), Xu constructed the following multi-objective programming model to obtain the priority vector:

$$\begin{aligned} &(\text{MOP1})\ \min\ \varepsilon_{ij} = \delta_{ij} \mid r_{ij} - 0.5(w_i - w_j + 1) \mid, \qquad i, j \in N \\ &\qquad s.t.\ \ w_i \geq 0, \quad i \in N, \ \sum_{i=1}^{n} w_i = 1 \end{aligned} \tag{2}$$

## 3 Priority Vector of Incomplete Fuzzy Preference Relation

Xu [26] asserted that an additive consistent incomplete fuzzy preference relation $R = (r_{ij})_{n\times n}$ satisfies (1). However, (1) does not hold for any additive consistent fuzzy preference relation. To show this, consider the following examples.

**Example 4.1.** For a decision-making problem, there are three alternatives under consideration. The decision maker provides his/her preferences over these three alternatives in the following fuzzy preference relation:

$$R_1 = \begin{bmatrix} 0.5 & 0.3 & 0.1 \\ 0.7 & 0.5 & 0.3 \\ 0.9 & 0.7 & 0.5 \end{bmatrix}.$$

$R_1$ is an additive consistent fuzzy preference relation. That is, all the known elements of $R_1$ satisfy the additive transitivity

$$r_{ij} = r_{ik} - r_{jk} + 0.5.$$

But for $R_1$, it is impossible to find $w = (w_1, w_2, w_3)^T$ such that

$$\delta_{ij} r_{ij} = \delta_{ij}[0.5(w_i - w_j + 1)], \qquad i, j \in N \text{, where } w_i \geq 0 \text{ and } \sum_{i=1}^{3} w_i = 1.$$

If we relax the constraint that $\sum_{i=1}^{3} w_i = 1$, we find that solution that satisfies

$$\delta_{ij} r_{ij} = \delta_{ij}[0.5(w_i - w_j + 1)], \qquad i, j \in N$$

would be $w_1 = c, w_2 = c + 0.4, w_3 = c + 0.8$, where $c$ is a nonnegative number. Since

$w_1 + w_2 + w_3 = 1.2 + 3c \geq 0$, it is impossible to find $w_i \geq 0$ such that $\sum_{i=1}^{n} w_i = 1$.

**Example 4.2.** Consider the following incomplete fuzzy preference relation of four alternatives:

$$R_2 = \begin{bmatrix} 0.5 & 0.375 & 0.25 & 0.125 \\ 0.625 & 0.5 & x & x \\ 0.75 & x & 0.5 & 0.375 \\ 0.875 & x & 0.625 & 0.5 \end{bmatrix}.$$

It is very easy to verify that all the known elements of $R_2$ satisfy the additive transitivity

$$r_{ij} = r_{ik} - r_{jk} + 0.5.$$

Following the Definition 2.4 $R_2$ is called an additive consistent incomplete fuzzy preference relation. We find that the priority vector $w = (w_1, w_2, w_3, w_4)^T$ that satisfies

$$\delta_{ij} r_{ij} = \delta_{ij}[0.5(w_i - w_j + 1)], \qquad i, j \in N$$

would be $w_1 = c, w_2 = c + 0.25, w_3 = c + 0.5, w_4 = c + 0.75$ , where $c$ is a nonnegative number. Since $w_1 + w_2 + w_3 + w_4 = 1.5 + 4c \geq 0$ , it is impossible to find $w_i \geq 0$ such that

$$\sum_{i=1}^{n} w_i = 1 \text{ and } \delta_{ij} r_{ij} = \delta_{ij}[0.5(w_i - w_j + 1)], \qquad i, j \in N .$$

## 4 Conclusion

We have proved the correspondence in equation (1) is invalid for additive consistent incomplete fuzzy preference relations which provides insight of the nature of additive consistency, i.e., the difficulty in prioritization assuming the weights are additive. To prioritize (incomplete) fuzzy preference relations, new correspondence between preference relation and priority vector has to be proposed.

## Acknowledgement

This research work was partially supported by the National Science Council of the Republic of China under grant No. NSC96-2416-H-019-004-MY2.

## References

1. Chiclana, F., Herrera, F., Herrera-Viedma, E.: Integrating three representation models in fuzzy multipurpose decision making based on fuzzy preference relations. Fuzzy Sets and Systems 97, 33–48 (1998)
2. Chiclana, F., Herrera, F., Herrera-Viedma, E.: Integrating multiplicative preference relations in a multipurpose decision-making model based on fuzzy preference relations. Fuzzy Sets and Systems 122, 277–291 (2001)
3. Chiclana, F., Herrera, F., Herrera-Viedma, F.E.: Reciprocity and consistency of fuzzy preference relations. In: De Baets, B., Fodor, J. (eds.) Principles of Fuzzy Preference Modelling and Decision Making, pp. 123–142. Academia Press (2003)
4. Chiclana, F., Herrera, F., Herrera-Viedma, F.E.: Rationality of induced ordered weighted operators based on the reliability of the source of information in group decision-making. Kybernetika 40, 121–142 (2004)
5. Cogger, K.O., Yu, P.L.: Eigenweight vectors and least-distance approximation for revealed preference in pairwise weight ratios. Journal of Optimization Theory and Application 46, 483–491 (1985)
6. Crawford, G., Williams, C.: A note on the analysis of subjective judgement matrices. Journal of Mathematical Psychology 29, 387–405 (1985)
7. Dubois, D., Prade, H.: Fuzzy Sets and Systems: Theory and Application. Academic Press, New York (1980)
8. Fan, Z.-P., Xial, S.-H., Hu, G.-F.: An optimization method for integrating tow kinds of preference information in group decision-making. Computers & Industrial Engineering 46, 329–335 (2004)

9. Herrera, F., Herrera-Viedma, E., Verdegay, J.L.: A sequential selection process in group decision-making with linguistic assessment. Information Sciences 85, 223–239 (1995)
10. Herrera, F., Martinez, L., Sanchez, P.J.: Managing non-homogeneous information in group decision making. European Journal of Operational Research 166, 115–132 (2005)
11. Herrera-Viedma, E., Herrera, F., Chiclana, F., Luque, M.: Some issues on consistency of fuzzy preference relations. European Journal of Operational Research 154, 98–109 (2004)
12. Jensen, R.E.: An alternative scaling method for priorities in hierarchical structures. Journal of Mathematical Psychology 28, 317–332 (1984)
13. Kacprzyk, J.: Group decision making with a fuzzy linguistic majority. Fuzzy Sets and Systems 18, 105–118 (1986)
14. Kacprzyk, J., Roubens, M.: Non-Conventional Preference Relations in Decision-Making. Springer, Berlin (1988)
15. Lee, H.-S.: Optimal consensus of fuzzy opinions under group decision making environment. Fuzzy Sets and Systems 132(3), 303–315 (2002)
16. Lee, H.-S.: On fuzzy preference relation in group decision making. International Journal of Computer Mathematics 82(2), 133–140 (2005)
17. Lee, H.-S.: A Fuzzy Method for Measuring Efficiency under Fuzzy Environment. In: Khosla, R., Howlett, R.J., Jain, L.C. (eds.) KES 2005. LNCS (LNAI), vol. 3682, pp. 343–349. Springer, Heidelberg (2005)
18. Lee, H.-S.: A Fuzzy Multi-Criteria Decision Making Model for the Selection of the Distribution Center. In: Wang, L., Chen, K., S. Ong, Y. (eds.) ICNC 2005. LNCS, vol. 3612, pp. 1290–1299. Springer, Heidelberg (2005)
19. Luce, R.D., Suppes, P.: Preference utility and subject probability. In: Luce, R.D., et al. (eds.) Handbook of Mathematical Psychology, vol. III, pp. 249–410. Wiley, New York (1965)
20. Orlovvsky, S.A.: Decision making with a fuzzy preference relation. Fuzzy Sets and Systems 1, 155–167 (1978)
21. Roubens, M.: Some properties of choice functions based on valued binary relations. European Journal of Operational Research 40, 309–321 (1989)
22. Saaty, T.L.: The Analytic Hierarchy Process. McGraw-Hill, New York (1980)
23. Tanino, T.: Fuzzy preference orderings in group decision-making. Fuzzy Sets and Systems 12, 117–131 (1984)
24. Xu, Z.S.: Generalized chi square method for the estimation of weights. Journal of Optimization Theory and Applications 107, 183–192 (2002)
25. Xu, Z.S.: Two methods for ranking alternatives in group decision-making with different preference information. Information: An International Journal 6, 389–394 (2003)
26. Xu, Z.S.: Goal programming models for obtaining the priority vector of incomplete fuzzy preference relation. International Journal of Approximate Reasoning 36, 261–270 (2004)
27. Xu, Z.S., Da, Q.L.: An approach to improving consistency of fuzzy preference matrix. Fuzzy Optimization and Decision Making 2, 3–12 (2003)
28. Zimmermann, H.J.: Fuzzy Set Theory and Its Applications. Kluwer, Dordrecht (1991)
29. Zhang, Q., Chen, J.C.H., He, Y.-Q., Ma, J., Zhou, D.-N.: Multiple attribute decision making: approach integrating subjective and objective information. International Journal of Manufacturing Technology and Management 5(4), 338–361 (2003)
30. Zhang, Q., Chen, J.C.H., Chong, P.P.: Decision consolidation: criteria weight determination using multiple preference formats. Decision Support Systems 38, 247–258 (2004)

# Fuzzy Multi-criteria Decision Making Based on Fuzzy Preference Relation

Hsuan-Shih Lee[1] and Chen-Huei Yeh[2]

[1] Department of Shipping and Transportation Management
Department of Computer Science
National Taiwan Ocean University
Keelung 202, Taiwan
Republic of China
[2] Yang Ming Marine Transport Corporation

**Abstract.** In this paper, we propose models to determine the weights of attributes by combining subjective information given by fuzzy preference relation and objective information given by the decision matrix. The fuzzy preference relations considered include the fuzzy preference relation with multiplicative transitivity and the fuzzy preference relation with additive transitivity. The proposed approaches are simpler than the previous methods that combine both subjective and objective information.

## 1 Introduction

Fuzzy preference relations have received a great deal of attention from researchers [1,3,4,5,7,8,9,10,11,12]. Fan et al. [2] investigated the multiple attribute decision making problem with fuzzy preference information on alternatives and proposed a decision aid approach to combine subjective fuzzy preference information with objective information for an overall assessment of the relative importance weights of the underlining attributes and select the best alternative. Approach proposed by Fan et al. has a greater intuitive appeal than the methods that consider only one type of information.

Consider a multiple criteria decision making problem with $n$ alternatives, $A_1, \dots, A_n$, and $m$ decision attributes (criteria), $C_1, \dots, C_m$. Each alternative is assessed with respect to each attribute. The assessment scores assigned to the attributes are the components of a decision matrix denoted by $X = (x_{ij})_{n \times m}$. Any incommensurability of the attributes is reconciled by normalizing the decision matrix $X = (x_{ij})_{n \times m}$. A common method of normalization is given as

$$b_{ij} = \frac{x_{ij} - x_j^{min}}{x_j^{max} - x_j^{min}}, i = 1, \dots, n, j \in \Omega_1 \tag{1}$$

$$b_{ij} = \frac{x_j^{max} - x_{ij}}{x_j^{max} - x_j^{min}}, i = 1, \dots, n, j \in \Omega_2 \tag{2}$$

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 980–985, 2008.

where $b_{ij}$ is the normalized attribute value, $x_j^{min} = \min_{1\leq i\leq n}\{x_{ij}\}$, $x_j^{max} = max_{1\leq i\leq n}\{x_{ij}\}$, and the set $\Omega_1$ and $\Omega_2$ are respectively the sets of benefit and cost attributes.

Let $B = (b_{ij})_{n\times m}$ be the normalized decision matrix and $W = (w_1, \ldots, w_m)^T$ the normalized vector of attribute weights such that $\sum_j^m w_j = 1$. The overall weighted assessment value of alternative $A_i$, $i = 1, \ldots, n$, is

$$d_i = \sum_{j=1}^{m} b_{ij} w_j, i = 1, \ldots, n, \tag{3}$$

where $b_{ij}$ represents objective information and $w_j$, $j = 1, \ldots, m$, are subjective weight variables. For brevity, (3) can be expressed in vector form as

$$D = BW \tag{4}$$

where $D = (d_1, \ldots, d_n)$ is a vector of the overall weighted assessment values for all the alternatives.

Suppose that fuzzy preference information on the alternatives provided by the decision maker is known and given in matrix form as

$$P = \begin{array}{c} \\ A_1 \\ A_2 \\ \vdots \\ A_n \end{array} \begin{array}{c} \begin{array}{cccc} A_1 & A_2 & \cdots & A_n \end{array} \\ \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & \cdots & \cdots \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{pmatrix} \end{array} \tag{5}$$

where $p_{ij}$ , $i = 1, \ldots, n$ and $j = 1, \ldots, n$ are membership grades characterized by the following membership function

$$p_{ij} = \mu_P(A_i, A_j) = \begin{cases} 1 & \text{if } A_i \text{ is definitely preferred to } A_j \\ c \in (0.5, 1) & \text{if } A_i \text{ is somewhat preferred to } A_j \\ 0.5 & \text{if there is no preference (i.e. indifference)} \\ f \in (0, 0.5) & \text{if } A_j \text{ is somewhat preferred to } A_i \\ 0 & \text{if } A_j \text{ is definitely preferred to } A_i \end{cases} \tag{6}$$

The matrix $P$ is called a fuzzy preference relation. Fan et al. assumed $P$ is multiplicative transitivity and employed the following quadratic programming problem to assess the attribute weights:

$$\begin{array}{ll} Min\ H(W) = \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} [p_{ij} \sum_{k=1}^{m} (b_{ik} + b_{jk}) w_k - \sum_{k=1}^{m} b_{ik} w_k]^2 \\ s.t. \quad \sum_{k=1}^{m} w_k = 1 \\ \qquad\ w_k \geq 0\ k = 1, \ldots, m. \end{array} \tag{7}$$

However, Wang et al. [6] showed that the solution of (7) provided by Fan et al. is not correct, which may have negative weights.

## 2 Multiplicative Transitivity

In this section, we are going to propose a linear programming model to obtain subjective weights $W = (w_1, \ldots, w_m)^T$ of attributes with respect to multiplicative transitivity.

Assume fuzzy preference relation

$$P = \begin{array}{c} \\ A_1 \\ A_2 \\ \vdots \\ A_n \end{array} \begin{array}{c} \begin{array}{cccc} A_1 & A_2 & \cdots & A_n \end{array} \\ \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & \cdots & \cdots \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{pmatrix} \end{array} \tag{8}$$

satisfies multiplicative transitivity. That is,

$$\frac{p_{ij}}{p_{ji}} \frac{p_{jk}}{p_{kj}} = \frac{p_{ik}}{p_{ki}} \text{ for } i, j, k = 1, \ldots, n. \tag{9}$$

Use $P$ to estimate $W$. Then

$$\begin{aligned} p_{ij} &= \tfrac{d_i}{d_i + d_j} \\ &= \tfrac{\sum_{k=1}^m b_{ik} w_k}{\sum_{k=1}^m (b_{ik} + b_{jk}) w_k} \\ \Rightarrow & \\ p_{ij} \sum_{k=1}^m (b_{ik} + b_{jk}) w_k &\approx \sum_{k=1}^m b_{ik} w_k \end{aligned} \tag{10}$$

By least deviation method, to estimate $W$ amounts to solve the following problem:

$$\begin{aligned} &Min\ H_1(W) = \textstyle\sum_{i=1}^n \sum_{j=1, j \neq i}^n |p_{ij} \sum_{k=1}^m (b_{ik} + b_{jk}) w_k - \sum_{k=1}^m b_{ik} w_k| \\ &s.t. \quad \textstyle\sum_{k=1}^m w_k = 1 \\ &\qquad\ \ w_k \geq 0\ k = 1, \ldots, m. \end{aligned} \tag{11}$$

(11) can be transformed into the following linear programming:

$$\begin{aligned} &Min\ G_1(W) = \textstyle\sum_{i=1}^n \sum_{j=1, j \neq i}^n (q_{ij} + r_{ij}) \\ &s.t. \quad \textstyle\sum_{k=1}^m w_k = 1 \\ &\qquad\ \ w_k \geq 0\ k = 1, \ldots, m \\ &\qquad\ \ p_{ij} \textstyle\sum_{k=1}^m (b_{ik} + b_{jk}) w_k - \sum_{k=1}^m b_{ik} w_k = q_{ij} - r_{ij},\ i, j = 1, \ldots, n \\ &\qquad\ \ q_{ij}, r_{ij} \geq 0. \end{aligned} \tag{12}$$

## 3 Additive Transitivity

Assume fuzzy preference relation

$$P = \begin{array}{c} \\ A_1 \\ A_2 \\ \vdots \\ A_n \end{array} \begin{array}{c} \begin{array}{cccc} A_1 & A_2 & \cdots & A_n \end{array} \\ \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & \cdots & \cdots \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{pmatrix} \end{array} \tag{13}$$

satisfies additive transitivity. That is,

$$p_{ik} = p_{ij} + p_{jk} \text{ for } i, j, k = 1, \ldots, n. \tag{14}$$

Use $P$ to estimate $W$. Then

$$\begin{aligned} p_{ij} &= \tfrac{d_i - d_j}{2} + 0.5 \\ &= \tfrac{\sum_{k=1}^{m} b_{ik} w_k - \sum_{k=1}^{m} b_{jk} w_k}{2} + 0.5 \end{aligned} \tag{15}$$

By least deviation method, to estimate $W$ amounts to solve the following problem:

$$\begin{aligned} &Min\ H_2(W) = \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} \left|2p_{ij} - \sum_{k=1}^{m} b_{ik} w_k + \sum_{k=1}^{m} b_{jk} w_k - 1\right| \\ &s.t. \quad \sum_{k=1}^{m} w_k = 1 \\ &\qquad\ \ w_k \geq 0\ k = 1, \ldots, m. \end{aligned} \tag{16}$$

(16) can be transformed into the following linear programming:

$$\begin{aligned} &Min\ G_2(W) = \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} (q_{ij} + r_{ij}) \\ &s.t. \quad \sum_{k=1}^{m} w_k = 1 \\ &\qquad\ \ w_k \geq 0\ k = 1, \ldots, m \\ &\qquad\ \ 2p_{ij} - \sum_{k=1}^{m} b_{ik} w_k + \sum_{k=1}^{m} b_{jk} w_k - 1 = q_{ij} - r_{ij},\ i, j = 1, \ldots, n \\ &\qquad\ \ q_{ij}, r_{ij} \geq 0. \end{aligned} \tag{17}$$

# 4 Illustration Examples

## 4.1 Example of Multiplicative Transitivity

The example used in Fan et al. [2] is adopted here. The example assumes a potential buyer intends to select a house from four alternatives ($S_1$, $S_2$, $S_3$ and $S_4$). When making a decision, the attributes considered include:

(1) $R_1$: house price
(2) $R_2$: dwelling area
(3) $R_3$: distance between every house and the work locality
(4) $R_4$: natural environment

Among four attributes, $R_2$ and $R_4$ are of benefit type, $R_1$ and $R_3$ are of cost type. The decision matrix with four attributes ($R_1$, $R_2$, $R_3$ and $R_4$) and four alternatives ($S_1$, $S_2$, $S_3$ and $S_4$) is presented as follows:

$$A = \begin{pmatrix} 3.0 & 100 & 10 & 7 \\ 2.5 & 80 & 8 & 5 \\ 1.8 & 50 & 20 & 11 \\ 2.2 & 70 & 12 & 9 \end{pmatrix} \tag{18}$$

which can be normalized into matrix $B$ as follows:

$$B = \begin{pmatrix} 0 & 1 & 5/6 & 1/3 \\ 5/12 & 3/5 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 2/3 & 2/5 & 2/3 & 2/3 \end{pmatrix} \tag{19}$$

Suppose the potential buyer gives his/her fuzzy preference relation on four alternatives satisfying multiplicative transitivity as follows:

$$P = \begin{pmatrix} 0.5 & 0.44 & 0.64 & 0.54 \\ 0.56 & 0.5 & 0.69 & 0.60 \\ 0.36 & 0.31 & 0.5 & 0.40 \\ 0.46 & 0.40 & 0.60 & 0.5 \end{pmatrix} \tag{20}$$

Using Eq. (12), we can obtain the weight vector on the attributes as follows:

$$W^* = (0.29760205, 0.132566132, 0.535431818, 0.0344)^T.$$

The ranking values of the four alternatives can be obtained by using Eq. (4), i.e.

$$d_1 = 0.59022598,\ d_2 = 0.738972351, d_3 = 0.33200205,\ d_4 = 0.631315698.$$

Thus, the ranking result of alternatives is

$$S_2 \succ S_4 \succ S_1 \succ S_3.$$

### 4.2 Example of Additive Transitivity

The sample problem is consider except that the fuzzy preference used is assumed to be of additive transitivity. The decision matrix with four attributes ($R_1$, $R_2$, $R_3$ and $R_4$) and four alternatives ($S_1$, $S_2$, $S_3$ and $S_4$) is presented in (18). The normalized matrix $B$ is shown in (19).

Suppose the potential buyer gives his/her fuzzy preference relation on four alternatives satisfying additive transitivity as follows:

$$P = \begin{pmatrix} 0.5 & 0.44 & 0.64 & 0.54 \\ 0.56 & 0.5 & 0.69 & 0.6 \\ 0.36 & 0.31 & 0.5 & 0.4 \\ 0.46 & 0.4 & 0.6 & 0.5 \end{pmatrix} \tag{21}$$

Using Eq. (17), we can obtain the weight vector on the attributes as follows:

$$W^* = (0.293374916, 0.216724825, 0.455500258, 0.0344)^T.$$

The ranking values of the four alternatives can be obtained by using Eq. (4), i.e.

$$d_1 = 0.607775041,\ d_2 = 0.707774702, d_3 = 0.327774916,\ d_4 = 0.60887338.$$

Thus, the ranking result of alternatives is

$$S_2 \succ S_4 \succ S_1 \succ S_3.$$

## 5 Conclusions

We have proposed linear programming models to determine the weights of attributes by combining subjective information given by fuzzy preference relation and objective information given by decision matrix. The fuzzy preference relations considered include the relations that satisfy multiplicative transitivity and additive transitivity. One of the possible future research directions is to deal with fuzzy preference relations that satisfy reciprocal property such as pairwise comparison matrices of AHP.

## Acknowledgement

This research work was partially supported by the National Science Council of the Republic of China under grant No. NSC95-2416-H-019-005-.

## References

1. Chiclana, F., Herrera, F., Herrera-Viedma, E.: Integrating multiplicative preference relations in a multipurpose decision-making model based on fuzzy preference relations. Fuzzy Sets and Systems 122, 277–291 (2001)
2. Fan, Z.-P., Ma, J., Zhang, Q.: An approach to multiple attribute decision making based on fuzzy preference information on alternatives. Fuzzy Sets and Systems 131, 101–106 (2002)
3. Herrera-Viedma, E., Herrera, F., Chiclana, F., Luque, M.: Some issues on consistency of fuzzy preference relations. European Journal of Operational Research 154, 98–109 (2004)
4. Orlovvsky, S.A.: Decision making with a fuzzy preference relation. Fuzzy Sets and Sytems 1, 155–167 (1978)
5. Tanino, T.: Fuzzy preference orderings in group decision-making. Fuzzy Sets and Systems 12, 117–131 (1984)
6. Wang, Y.-M., Parkan, C.: Multiple attribute decision making based on fuzzy preference information on alternatives: Ranking and weighting. Fuzzy Sets and Systems 153, 331–346 (2005)
7. Xu, Z.S.: Two methods for ranking alternatives in group decision-making with different preference information. Information: An International Journal 6, 389–394 (2003)
8. Xu, Z.S.: On compatibility of interval fuzzy preference matrices. Fuzzy Optimization and Decision Making 3, 217–225 (2004)
9. Xu, Z.S.: Goal programming models for obtaining the priority vector of incomplete fuzzy preference relation. International Journal of Approximate Reasoning 36, 261–270 (2004)
10. Xu, Z.S.: On method for uncertain multiple attribute decision making problems with uncertain multiplicative preference information on alternatives. Fuzzy Optimization and Decision Making 4, 131–139 (2005)
11. Xu, Z.S.: A procedure for decision making based on incomplete fuzzy preference relation. Fuzzy Optimization and Decision Making 4, 175–189 (2005)
12. Zhang, Q., Chen, J.C.H., He, Y.-Q., Ma, J., Zhou, D.-N.: Multiple attribute decision making: approach integrating subjective and objective information. International Journal of Manufacturing Technology and Management 5(4), 338–361 (2003)

# The Development of the Financial Learning Tool through Business Game

Yasuo Yamashita[1], Hiroshi Takahashi[2], and Takao Terano[3]

[1] Department of Computational Intelligence and Systems Science, Tokyo Institute of Technology, Japan
[2] Graduate School of Business Administration, Keio University, Japan
[3] Department of Computational Intelligence and Systems Science, Tokyo Institute of Technology, Japan

**Abstract.** Recently financial education has become more important because financial technology is highly developing and financial market is growing in importance. In this research, we apply Business Game method to financial education. Especially we focused on learning of asset allocation. As a result of intensive experiments, we found that (1) players learned not to take excessive risk through business game and (2) they recognize the importance of risk control by our experiments. These findings indicate that our approach is valid for financial education.

**Keywords:** business game, WWW browser, WWW server, investment decision making.

## 1 Introduction

In recent years, the investment in financial assets is attracting the interest by making the rapid spread of financial products with a risk, a rise of concern about public and private pension asset management, etc. into a background. Risk management especially plays an important role in the investment in financial assets, and research is eagerly done to a market risk, a credit risk, etc. also in the present.

Although much technique is proposed, a mean-variance model is mentioned as technique for which the risk management technique is most widely used[1][2][3]. There are many models using an advanced formula in the model generally proposed in the finance. It is pointed out that it is not necessarily easy to understand the application method to a practical concrete investment move.

Therefore, in order to promote rational decision making in an actual investment, there is the need of performing training for performing suitable decision making in the situation where the feature of an actual investment was adopted.

From these backgrounds, the need of learning the investment in financial assets in the situation where the element of the investment environment in an actual market was taken in, is high. In this research, it tries to apply the business game technique to financial education. Business game is a game of the human participatory type using a computer system. For example, in the field of business

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 986–993, 2008.

administration, business game is used as the technique of learning a business model effectively[1].

It aims at showing the technique of learning the investment in effective financial assets using the framework of business game in this research. In this research, first, the environment which took in the element of an actual investment was built, next it experimented by using an actual human being as a player. In the following chapter, the model used for analysis is explained and a result is shown in Chap.3. Chap.4 is a conclusion.

## 2 Method

### 2.1 System of Business Game

As environment required for development of the system in this research, and execution of an experiment, the business game development tool of Graduate School of System Management( henceforth, GSSM) was used[4]. The development tool of GSSM is constituted by Business Model Description Language (BMDL) and Business Model Development System (BMDS).[2] By describing the source code of BMDL which is a simple programming description language, an HTML file, a CGI file, etc. the object for game managers (facilitator) and for game users (player) can be created by BMDS.

### 2.2 Model of Business Game

Some processes exist in decision making of investment. In this experiment, the business game which focused on asset allocation was built. Asset allocation is said to decide about 90 percent[3] of investment performance by the success or failure, and is important decision making in asset management[5]. In this analysis, a player opts for the asset allocation of 4 assets (a domestic bond, a domestic stock, a foreign bond, a foreign stock) as a money manager of financial assets every six months.

Some rating scales are proposed about the performance evaluation after investing. A Sharpe ratio exists in the index most widely used in financial business[6]. This index can be computed like ”Sharpe ratio” = ”average of monthly earning rate” / ”standard deviation of a monthly earning rate.” In this experiment, each player is given raising a Sharpe ratio as much as possible as a purpose of business game. Moreover, in management of funds including an actual investment fund etc., aggravation of the performance of investment will generate cancellation of a fund,

[1] Especially, the system of University of Tsukuba which can perform under the Web environment is applied to various fields from the height of the effectiveness.

[2] BMDS is changed and used in part so that the experiment of this research may be possible.

[3] The remaining ten percent is the sector allocation effect and the individual issue selective effect. In the experiment of this research, each asset should use the index and it is assumed that neither sector allocation nor individual issue selection is made. I would like to make these detailed analysis into a future subject.

etc. The situation of such an actual market shall be made into a background, and, as for the player below the Sharpe ratio average of all the players, a working capital shall be canceled in this experiment at the time of the end of a game.

In the start of the round in every six months, the player can obtain the reference information. Next, a player makes decisions for every round based on these information (input of an asset allocation ratio).[4] The player can see the output information in the end of each round.[5] An experiment shall be ended in the place which repeated such decision making, and performed and carried out 9 round continuation.

Moreover, to the player, the ranking between the players on the basis of a Sharpe ratio is given as output information in the end of each round. This is a device for building a competitive situation and urging the participation to an earnest experiment of a player in order to check the existence of influence in decision making of the player of precedence information.

Business game of the above contents was carried out 3 times. A player is the Okayama University department-of-economics student (a second grader, a third grader), and is a beginner mostly about finance theory. In the 1st time and the 2nd experiment, it carried out into four groups in every several persons, and the 3rd experiment was conducted in each individual (eight persons).

The time for carrying out explanation of introduction and an experiment and decision making of a round 1 was taken about 30 minutes. The procedure of repeating decision making of each round every about 10 minutes after that performed. One experiment took about about 3 hours.

Moreover, in order to investigate a student's learning effect, it experimented in the experiment 1 which uses a student as a player to affiliation member 4 person (primary financial analyst test success level grade) and the setup in an institutional investor's investment section.

## 3 Result

### 3.1 The Feature of an Investment Move

Fig.1 shows transition (p1-p4 in a figure) of the rate of return for the past domestic stock six months (A2 in a figure), and the domestic stock ratio of each player in the start of each round in experiment 1. The player 2 can check that the investment ratio to domestic stocks is increasing and decreasing according to a rise and descent of the latest six-month rate of return of domestic stocks from Fig.1. Change of the domestic stock ratio of this player and the correlation coefficient of the domestic per share earnings ratio were 0.83, and were a high value.

Table 1 is the result of calculating correlation with the rate of return of each property, and the asset allocation ratio of each player. It can check that some

[4] The sum total of a ratio is 100% and the ratio of each asset is made into 100% or less of range 0% or more.

[5] In order to bring close to actual asset management environment, rate of return and a Sharpe ratio are calculated based on the actual asset price of a period with the past. However, a player is not told when a period is.

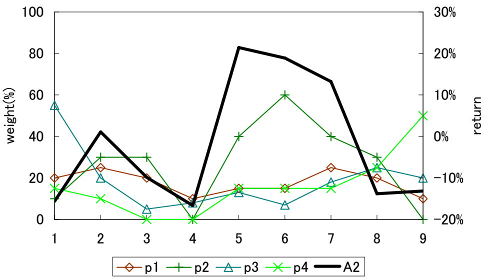


**Fig. 1.** Return of a domestic stock (A2: right axis) and domestic stock ratio in experiment 1 (p1-p4) Transition

**Table 1.** Correlation with rate of return on assets and asset allocation ratios

| exp. | group (player) | dome. bond | dome. stock | foreign bond | foreign stock |
|---|---|---|---|---|---|
| 1 | 1 | -0.38 | 0.19 | 0.28 | -0.39 |
| | 2 | -0.10 | **0.83** | 0.38 | -0.23 |
| | 3 | 0.21 | -0.37 | -0.16 | -0.18 |
| | 4 | **0.55** | -0.10 | 0.25 | **0.80** |
| 2 | 1 | -0.03 | -0.06 | -0.08 | -0.04 |
| | 2 | -0.11 | -0.41 | -0.52 | 0.44 |
| | 3 | -0.55 | 0.34 | 0.27 | -0.68 |
| | 4 | -0.55 | -0.18 | **0.59** | 0.18 |
| 3 | a | 0.35 | 0.40 | -0.50 | 0.36 |
| | b | -0.21 | 0.10 | 0.16 | 0.32 |
| | c | -0.23 | **0.63** | -0.02 | -0.24 |
| | d | 0.20 | -0.24 | 0.26 | 0.15 |
| | e | -0.09 | **0.80** | **0.72** | **0.61** |
| | f | 0.03 | **0.62** | -0.10 | **0.62** |
| | g | **0.76** | 0.33 | 0.44 | -0.29 |
| | h | 0.12 | **0.55** | **0.65** | -0.47 |

which show high correlation are between change of the investment ratio to each property, and rate of return from Table 1.

These results suggest a possibility of having taken the investment move as a positive feedback trader with same player[7]. Like an actual market, among individual investors with restrictive information and analysis ability, there are someone who take the investment move which was similar with the positive feedback trader.

### 3.2 Reduction of Excessive Risk Acquisition Action

Fig.2 and Fig.3 show risk transition of each group for every round from experiment 1 to experiment 3.[6] "1" and "2" are the amounts of risks of experiment 1 and experiment 2 among a figure, respectively, and "3(X)" (X is the character

[6] Risks are the standard deviation calculated in the experiment 1 and the experiment 2 by the covariance by the monthly rate of return at the time of a round start (120 months), and the standard deviation calculated by the covariance by monthly rate of return (60 months) in the experiment 3.

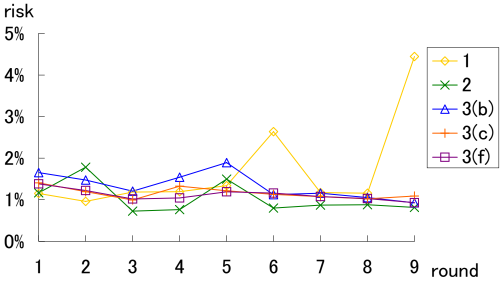


**Fig. 2.** Risk transition of group 2

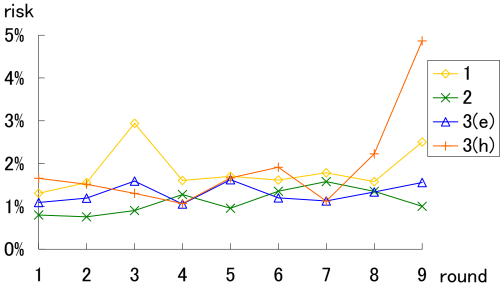


**Fig. 3.** Risk transition of group 4

from ”a” to ”h”) is a risk of the player X in experiment 3. Since the experiment 3 was conducted not in a group but in the individual, risk transition for every individual of a group and experiment 3 which belonged in the experiment 1 and the experiment 2 is drawn all over the same figure.

If these figures are seen, it can check that the group which is carrying out decision making which took the extremely big risk exists. For example, in Fig.2, a group 2 can check having taken the extremely high risk at the round 6 and the round 9 in experiment 1.

Moreover, in Fig.3, a group 4 can check having taken the high risk at the round 3 in experiment 1. Thus, decision making which takes an extreme risk has been checked in the experiment which used the student as the player. In finance theory, it is hard to consider that decreasing a risk by diversified investment is shown and the employment person who is the meaning and by whom it was trained by the institutional investor etc. takes such an extreme risk. So, in this experiment, in order to compare with a student's result, the same experiment was conducted for the affiliation member in an actual institutional investor's investment section.

Fig.4 shows risk transition of the experiment which used the institutional investor affiliation member 4 person as the player by the same setup as the experiment 1 in the case of a student. In this case, it can check not having carried out decision making which took the extremely high risk as compared with the result in the case of a student. As a reason which these differences produce, it is

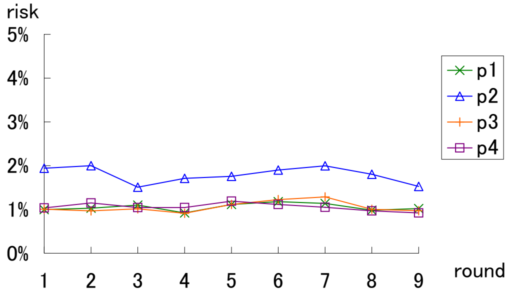


**Fig. 4.** Risk transition of institutional investor affiliation member

**Table 2.** Standard deviation of a risk:unit %

| group | exp.1 | exp.2 | exp.3 | | | investor member |
|---|---|---|---|---|---|---|
| 1 | 0.21 | 1.00 | 0.18(d) | 0.29(g) | – | 0.08 |
| 2 | 1.14 | 0.37 | 0.32(b) | 0.14(c) | 0.14(f) | 0.19 |
| 3 | 0.50 | 0.45 | 0.23(a) | – | – | 0.13 |
| 4 | 0.53 | 0.29 | 0.23(e) | 1.16(h) | – | 0.08 |
| median | 0.51 | 0.41 | 0.23 | | | 0.11 |
| diff. of a median | – | 0.10* | 0.28** | | | 0.40** |

*Significant with the 5% level,
**Significant with the 1% level. Median test(one-side)
The difference of a median expresses the diff. of exp. 1 and exp.2 with exp.2.
The diff. of exp.2 and exp.3 is expressed with exp.3.
The diff. of the median of exp.1 and an institutional investor affiliation member is expressed with an institutional investor affiliation member.
An inner character at a parenthesis is a character which distinguishes the player in exp.3.

mentioned as one factor that the student is not learning the importance of risk management compared with an institutional investor affiliation member.

Although the experiment to a student was conducted 3 times, on the whole it can check that it is in the tendency whose investment move which takes an extreme risk decreases as it passes through an experiment. For example, although the player had taken the extremely high risk at the round 6 or the round 9 in the experiment 1 in Fig.2, such an extremely high risk is not taken in experiment 2.

Table 3.2 shows the standard deviation of the risk of experiment 1, experiment 2, experiment 3, and an institutional investor affiliation member. The standard deviation of a risk here is what calculated the standard deviation of the risk of all the rounds for every player, and expresses the variation in a risk. If this value is high, it can be considered that the extreme risk is taken.

When the standard deviation of the risk of the player 2 in Table 3.2 and a player 3 is seen, it turns out that the investment move which takes a extreme risk is decreasing as an experiment is repeated. Although the player 1 had taken the extremely high risk at a round 1 in the experiment 2, it will not take an in general extreme risk in experiment 3. A player 4 will not take a extremeer risk, if a player 8 removes having taken the extremely high risk in experiment 3. If the median of the standard deviation of a risk is seen for every experiment in order to grasp the whole tendency, since there are few samples, it can check falling with 0.51%, 0.41%, and 0.23% as an experiment is repeated. Even if it sees the result of a median test, a significant difference is observed in the median of experiment 1 and experiment 2 with the level 5%. Moreover, a significant difference is looked at by the median of experiment 2 and experiment 3 with the level 1%.

In the actual market, the institutional investor's fund manager is performing competition with other fund managers through the performance of the fund which oneself takes charge of. This experiment is also conducted on a setup which can check the precedence information of other players reflecting the situation of such an actual market.

### 3.3 Moral Hazard in End of Experiment

If it analyzes in detail about Fig.2 and Fig.3 in the foregoing paragraph, in the end of an experiment, it can check that the risk is going up extremely. It is tended in the whole player to set especially a phenomenon such to a player with low ranking. Such a phenomenon is the action which the bad investor of performance is considered to have produced in order to apply an all-or-nothing great victory negative in the end of an experiment, and can say also with a moral hazard in the meaning[8].

The model in this experiment can be regarded as a principal agent model in a game theory, if a working capital truster is made principal and it makes the fund manager (player) an agent. Although the result of the player which is an agent is observed as investment performance in an experiment, a risk level of a player choosing is a setup which is not manageable from a principal employment truster. Therefore, it can be said that what actually starts the moral hazard which is going to raise a result in the end of an experiment into the low rank player which was conscious of competition even if it takes a surplus risk appeared.

Such a moral hazard is a phenomenon seen also in an actual market. The hedge fund etc. in which employment performance got worse apply a leverage, an excessive risk is taken, and there is an example of making invested assets damage on the contrary. That such a moral hazard was actually able to be seen shows that this experiment can be reproducing actual operational environment, and it is an interesting result.

## 4 Conclusion

This research showed that decision making of investment to financial assets could be learned with the business game technique.

First, in the player, it was shown that an investment move like a positive feedback trader may be taken. It is an interesting result that investor action which is seen in an actual market is reproduced also in this experiment.

Next, decision making of investment showed that a player learned that it is important to make it not take an extreme risk. The technique of this research proposes the new approach to the field considered to be difficult so far "study of the financial investment in a competitive position." It is meaningful from a viewpoint of financial education.

The last, we showed a phenomenon, the moral hazard taking the risk by the low player of ranking is extremely high in final stage of an experiment. The fact the phenomenon has been observed shows that it can say we reproduce actual operational environment, and it proves the validity of this technique.

## References

1. Markowitz, H.M.: Portfolio selection. Journal of Finance 7, 77–91 (1952)
2. Tobin, J.: Liquidity preference as behavior toward risk. Review of Economic Studies 26, 65–86 (1958)
3. Sharpe, W.F.: Portfolio analysis. Journal of Financial and Quantitative Analysis 2, 76–84 (1967)
4. Terano, T., Suzuki, H., Kuno, Y., Fujimori, H., Shirai, H., Nishio, H., Ogura, N., Takahashi, M.: Understanding your business through home-maid simulatior development. Developments in Business Simulation and Experimential Learning (26), 65–71 (1999)
5. Brinson, G.P., Singer, B.D., Beebower, G.L.: Determinants of portfolio performance ii: An update. Financial Analysts Journal 47, 40–48 (1991)
6. Sharpe, W.F.: Mutual fund performance. Journal of Business 39, 119–138 (1966)
7. DeLong, J.B., Shleifer, A., Summers, L., Waldmann, R.: Positive feedback investment strategies and destabilizing rational speculation. Journal of Finance 45, 375–395 (1990)
8. Rasmusen, E.: Games and Information: an introduction to game theory, 4th edn. Blackwell Publishing Ltd, Malden (2007)

# A Method for Ensuring Consistency of Software Design Information in Retrospective Computer Validation

Masakazu Takahashi[1,2], Satoru Takahashi[3], and Yoshikatsu Fujita[4]

[1] University of Yamanashi, Takeda 4-3-11, Kofu, Yamanashi, Japan
mtakahashi@yamanashi.ac.jp
[2] Galaxy Express Corporation, 18-16-1 Hamamatsucho, Minato-ku, Tokyo, Japan
masakazu_takahashi@galaxy-express.co.jp
[3] Mitsui Asset Trust Banking Co., Ltd, 2-1jyonan-mishima, Tokyo, Japan
satoru_1_Takahashi@mitsuitrust.co.jp
[4] Graduate School of Business Sciences, University of Tsukuba, Otsuka3-29-1, Bunkyo, Tokyo, Japan
fujita.yoshikatsu@jp.panasonic.com

**Abstract.** This paper proposes a method for ensuring consistency of software design information when we conduct Retrospective Computer Validation (RCV). RCV is proof tasks that quality of in-service software related to drug manufacturing (DMSW) is adequate. When we conduct RCV in first time, we show design adequacy of DMSW to collect existing documents and running records. When we remodel DMSW, we show design adequacy to develop new documents and conduct enough tests. It is difficult to ensure consistency because there are some differences of description items and fineness between existing and new documents. This problem made RCV enforcement difficult. We will solve this problem to conduct following countermeasures; 1) define document architecture, 2) define description items and contents in every document, and 3) implement database that manages description items and relationships between documents. Consequently, we can clarify affected range, when some remodeling occurs. To modify design information in affected range, we can maintain adequacy of design information of DMSW. As the result, we can conduct RCV smoothly.

**Keywords:** Retrospective Computer Validation, Pharmaceutical Production System, Consistency, Design Information Database.

## 1 Introduction

This paper proposes a method ensuring consistency of design information of in-service Drug Manufacturing Software (DMSW) between development documents in Retrospective Computer Validation.

In these days, many troubles related to drug qualities occurred. Regulatory authorities of drug manufacturing required to validate that drugs with adequate qualities were manufactured for Drug Manufacturing Companies (DMCs). DMSW that is used for

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 994–1001, 2008.

facility control and quality data management is required to validate that it works adequately according to drug manufacturing procedure. DMCs continue to validate that DMSW is working adequately from operation start to operation completion. This task is called as Computer Validation (CV).

CV method that is applied to in-service DMSW is called as Retrospective Computer Validation (RCV). RCV was defined as a validation method for DMSW that had been used before requiring CV by regulatory authorities. RCV is an indirect validation method, because working DMSW correctly is validated based on the evidences of existing development documents and operation records (hereafter, IRCV: Initial RCV).

In the actual operation, some software modifications and Additional RCV (ARCV) are required according to some modification about manufacturing facilities and processes. In those cases, it occurs following problems related to development documents that are collected in IRCV and created in ARCV occurs.

Problem 1: Differences of varieties of development documents.
Problem 2: Differences of description items and levels.
Problem 3: Laxness of correspondences between description items.

As a result, DMCs are pointed out by regulatory authorities that it cannot validate the adequate behavior of DMSW, and this makes drug manufacturing difficult.

Researches for solving Problem 1 are developments of efficient CV procedures. CV procedures for personal computer [8] and programmable logic controller [9] based DMSW are proposed, but those methods do not focus on RCV. Researches for solving Problem 2 are usage of specification description language and development documents creation method using patterns. There are Z language [7] and VDM-SL [3] etc. as a specification description language, and those describes specifications using mathematical method. In methods for development document creation using pattern, some templates of development documents are prepared as patterns, and specified development documents are created reusing such templates [4], [10]. Researches for solving Problem 3 are researches about consistency and traceability. A research about consistency is a method that consistencies between specifications and software are verified formally using specification description language [5]. Comparison and evaluation about traceability approaches exists conducted by Siti [6].

It is difficult to apply those researches to RCV, because subject is different, advanced technologies are required, and over-head of tasks are increased.

## 2 Current Status of Retrospective Computer Validation

At first, Table 1 shows software subject to RCV. As the result, we find that software subject to RCV is written in procedural programming languages, such as C, BASIC and Macro language, is small-scale, and has simple structure.

Next, Figure 1 shows RCV procedure [2]. There is no definition of development documents architecture, description item and description level about RCV. Consequently, styles of development documents prepared by DMCs in IRCV are various. DMCs tend to create minimum development documents or not to create development documents to reduce costs related to RCV.

In the other hand, when conducting ARCV, development documents architecture, description items and description levels are defined, because development documents are created according to the standard CV regulation. This is called as "Good Automated Manufacturing Practice (GAMP) [1]."

As a result, when we conduct ARCV using development documents prepared in IRCV, we cannot secure consistency between development documents of existing part and newly creating part. This causes problems discussed in section 1.

**Table 1.** Software Variations Subject to Retrospective Computer Validation

| Intended Purpose | Data Management | Report Development | Technical Computing | Facility Control |
|---|---|---|---|---|
| Programming Language | BASIC,<br>Macro | BASIC,<br>C | BASIC,<br>FORTRAN,<br>Macro | BASIC,<br>C |
| Structure | Simple | Simple | Simple | Simple |
| Algorism | Simple | Simple | Simple | Rather Complex |
| Size [lines of codes] | 500 - 700 | 200 – 300 | 200 - 400 | 700 - 1000 |

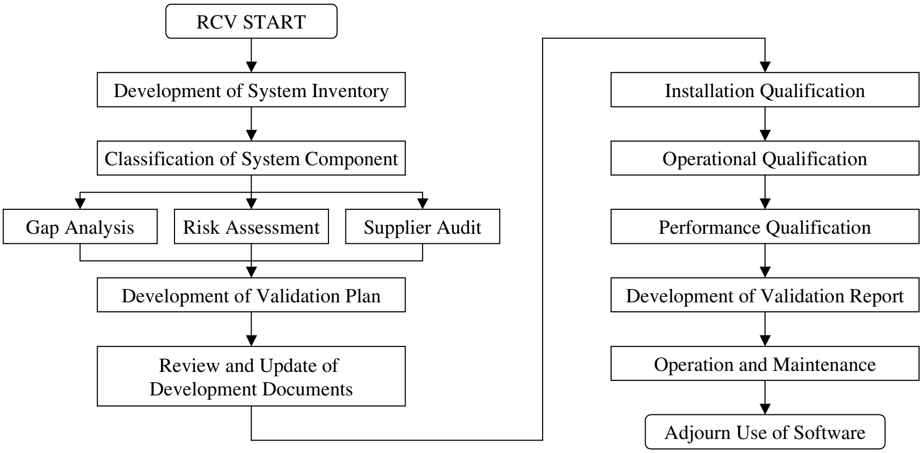

RCV START
Development of System Inventory
Classification of System Component
Gap Analysis
Risk Assessment
Supplier Audit
Development of Validation Plan
Review and Update of Development Documents
Installation Qualification
Operational Qualification
Performance Qualification
Development of Validation Report
Operation and Maintenance
Adjourn Use of Software

**Fig. 1.** Procedure of RCV

## 3 Outline of the Proposed RCV

To solve problems sated in section 1, following countermeasures are necessary.

Countermeasure 1: Define development documents architecture.
Countermeasure 2: Define description items and levels in development documents.
Countermeasure 3: Clarify relations between description items in development documents.

At first, we discuss the Countermeasure 1. It is necessary to prepare development documents considering ARCV when conducting IRCV. It is best way to prepare development documents and to wrie description items required by GAMP's CV regulation. Figure 2 shows the development documents architecture in the proposed method.

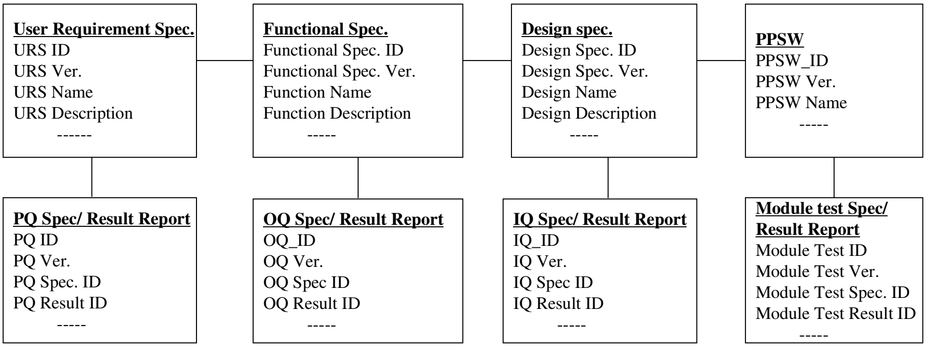


**Fig. 2.** Development Documents Architecture

Next, we discuss the Countermeasure 2. We reversely create above development documents from in-service software targeted at RCV. We prepare templates of development documents, such as Design Specification (DS), Functional Specification (FS), and User Requirement Specification (URS) according to GAMP's CV regulation. Items in each Table of Figure 2 show description items. We create development documents filling out items in those templates. We utilize Structure Chart (SC) and Flow Chart (FC) to fill out description items. The reasons why we use SC and FC are easy to reversely create from in-service software, are easy to describe items without ambiguity and are easy to modify affected part of development documents related to software modification in ARCV.

Figure 3 shows the procedure reversely creating SC and FC from in-service DMSW. FC is created by analyzing in-service software. Combined subroutines and functions are consolidated as modules. The execution sequence of modules is defined as SC. The creation method of development documents using FC and SC is as followings.

**[Design Specification]**

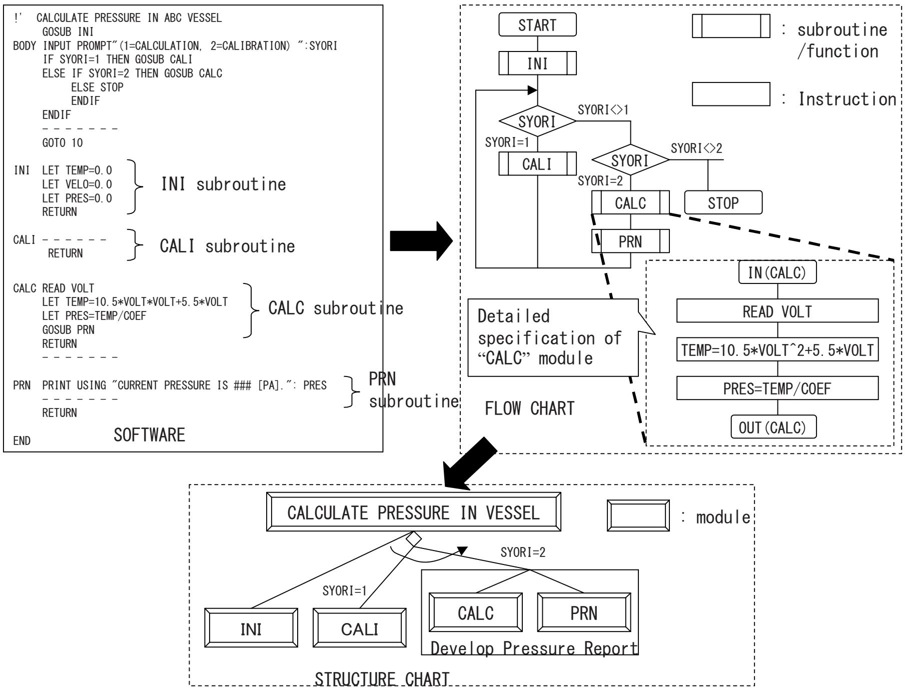
DS in subroutine and function unit is created by giving adequate ID, Version, DS name and embedding FC into Design Description of the DS template. All DSs in subroutine unit are collected and whole DS is created.

**[Functional Specification]**

FS is created by combining some DSs in function unit, giving adequate ID, version and name, and embedding summary of single function into Function Description of FS template. All FSs are collected and whole FS is created.

**[User Requirement Specification]**

URS is created by combining some FSs in procedure unit, giving qdequate ID, version and name, and embedding summary of single operation of URS. All URSs are collected and whole URS is created.



```
!'  CALCULATE PRESSURE IN ABC VESSEL
     GOSUB INI
BODY INPUT PROMPT"(1=CALCULATION, 2=CALIBRATION) ":SYORI
     IF SYORI=1 THEN GOSUB CALI
     ELSE IF SYORI=2 THEN GOSUB CALC
          ELSE STOP
          ENDIF
     ENDIF
     - - - - - - - -
     GOTO 10

INI  LET TEMP=0.0
     LET VELO=0.0
     LET PRES=0.0
     RETURN

CALI - - - - - - -
      RETURN

CALC READ VOLT
     LET TEMP=10.5*VOLT*VOLT+5.5*VOLT
     LET PRES=TEMP/COEF
     GOSUB PRN
     RETURN
     - - - - - - - -

PRN  PRINT USING "CURRENT PRESSURE IS ### [PA].": PRES
     - - - - - - - -
     RETURN

END
```

**Fig. 3.** Creation Method of Development Documents

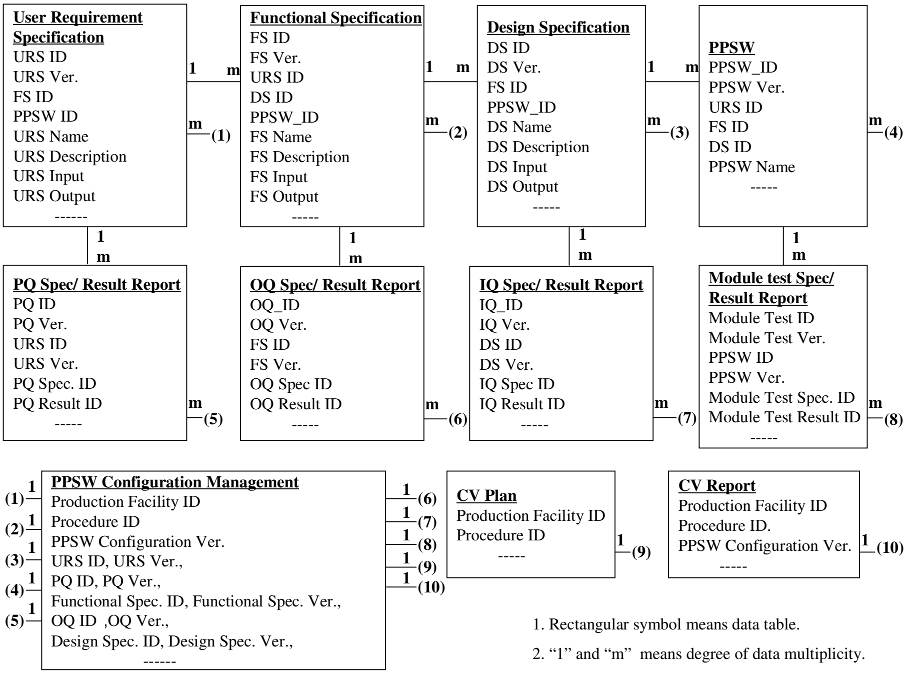
At Last, we discuss the countermeasure 3. To trace design information between documents, it is necessary to define relations between URS, FS, and US. Furthermore, it is necessary to clarify relations between test, IQ, OQ, and PQ specifications in ARCV. Figure 4 shows database structure to manage design information concerning RCV.

## 4 Evaluation of Proposed RCV Method

We evaluate proposed Countermeasures in this section.

At First, we evaluate the Countermeasure 1. The development document architecture was defined. This made no differences of development documents between created in IRCV and ARCV. And the Architecture follows GAMP's CV regulation. There is no problem that the architecture is applied to RCV. Consequently, we can solve Problem 1, such as "Differences of varieties of development documents" between in IRCV and ARCV, by applying Countermeasure 1.

At second, we evaluate the Countermeasure 2. The description items and description levels of development documents are defined in Solution 1. The description items follow GAMP's CV regulation. The description levels use FCs, SCs and summaries reverse-created from in-service DMSW. In DS, GAMP requires to describe subroutine specifications that are development units of DMSW. This satisfies



1. Rectangular symbol means data table.

2. “1” and “m” means degree of data multiplicity.

**Fig. 4.** Outline of Design Information Management Database

GAMP’s requirements, because FC corresponds to every subroutine’s specification. In FS, GAMP requires to describe functions. To describe function using SC’s modules satisfies GAMP’s requirement, because a SC’s module, as a unit of SC, correspond to a function combining some subroutines. In URS, GAMP requires to describe DMSW’s requirements. Whole SC corresponds to procedure that is made combining several functions. To describe requirements using whole SC satisfies GAMP’s requirements, because whole SC corresponds to DMSW’s operation requirements on the user’s viewpoint. As a result, we can solve Problem 2, such as “Differences of description items and description levels” between in IRCV and ARCV, by applying Countermeasure 2.

At last, we evaluate the Countermeasure 3. The consistencies of DMSW’s design information described in development documents were secured by using database in the Solution 3. We evaluate securing design information in DS, FS, and URS level.

(1) Modification in DS (FC) level

In this case, we have to modify DS ID, DS Version, DS Name, and DS Description described in related DS table that DS input corresponds to DS output in modified DS table. This secures consistencies between related DS tables.

(2) Modification in FS (SC’s module level)

In this case, we have to modify FS ID, FS Version, FS Name, and FS Description described on related FS table that FS input corresponds to FS output in modified FS table. This secures consistencies between related FS tables. Additionally, as same procedure of (1), DS tables that are pointed out by DS ID in the FS Table are modified. This secures consistencies between related DS tables. As a result, consistencies between related tables can be secured.

(3) Modification in URS (whole SC level)

In this case, we have to modify URS ID, URS version, URS Name and URS Description described on related URS tables that URS input corresponds to URS output in modified URS table. This secures consistencies between related URS tables. Additionally, as same as procedure (2), FS tables that are pointed out by FS ID in URS Table are modified. This secures consistencies between related FS tables. And DS tables that are pointed out by DS ID in the FS table are modified. This secures consistencies between related DS tables. As a result, we can secure consistencies between all related tables when we conduct URS level modification.

As a result, we can solve Problem 3, such as “Laxness of correspondences between description items” between in IRCV and ARCV, using Countermeasure 3.

## 5 Conclusion

We confirmed that consistencies of in-service DMSW’s design information can be secured using the proposed method. This makes that DMC can conduct adequate RCV. We consider that the proposed method contributes to manufacture drugs that have adequate quality, and DMC’s competitive edge will be improved.

We will develop methods that can identify ranges of development documents that have to be modified when DMSW is modified and generate test specification automatically related to modified software. These will contribute to conduct adequate RCV efficiently.

## Acknowledgement

A method for ensuring consistency between development documents was researched in "Sophisticated Mission Integration Design Technology Development for Next Generation Transportation System Basic Technologies Development Project" from NEDO (New Energy and Industrial Organization's 2007 operation). This paper reported that the research results were applied to Retrospective Computer Validation and were confirmed usefulness. I would like to take this opportunity to express thanks.

## References

[1] GAMP forum: GAMP 4 Guide for Validation of Automated System in Pharmaceutical Manufacture. International Society for Pharmaceutical Engineering (2001)

[2] Hagiwara, K., et al.: Computer Validation, FDA, Ministry of Welfare ER/ES Guidelines and Applications, Jyouhou Kikou (2006) (in Japanese)

[3] Middelburg, C.A.: Logic and Specification: Extending Vdm-Sl for Advanced Formal Specification. Chapman & Hall, Boca Raton (1993)

[4] Satou, M., Inoue, M., Inoue, T., Yamada, T.: A Proposal of OSS Requirement Definition Using Patterns in Equipment and NW Management Domain. IEOCE Technical Reserch Report 105, 85–94 (2005) (in Japanese)

[5] Satyananda, T.K., Lee, D., Kang, S.: Formal Verification of Consistency between Feature Model and Software Architecture in Software Product Line. In: Proc. of ICSEA 2007, pp. N/A (2007)

[6] Siti, R., Wan, M.N., Wan, K., Abdul, H.A.: An Evaluation of Traceability Approaches to Support Software Evolution. In: Proc. of ICSEA 2007, pp. N/A (2007)

[7] Spivey, J.M., Understanding, Z.: A Specification Language and Its Formal Semantics. Cambridge Univ. Press, Cambridge (2008)

[8] Takahashi, M., Tsuda, K.: An Efficient Computer Validation Method for Pharmaceutical Facility Motion Control Software. Journal of Information Processing Society of Japan 45(12), 2869–2879 (2004) (in Japanese)

[9] Takahashi, M.: An Efficient Method for Computer Validation of Pharmaceutical Production Facility Control Software Using Programmable Logic Controller. Transactions of The Society of Instrument and Control Engineers 42(8), 949–958 (2006) (in Japanese)

[10] Zeida, R., et al.: Automatic Conversion from Specifications written in Japanese to Class Diagrams of UML, Technical Report of The Institute of Electronics. Information and Communication Engineers, Pattern Recognition and Media Understanding, Vol.105, No. 615, pp. 159–164 (2005)

# Egocentrism Presumption Method with N-Gram for e-Business

Nobuo Suzuki[1] and Kazuhiko Tsuda[2]

[1] KDDI Corporation
Iidabashi 3-10-10, Chiyoda, Tokyo 102-8460, Japan
nu-suzuki@kddi.com
[2] Graduate School of Buisiness Sciences, University of Tsukuba
Otsuka 3-29-1, Bunkyo, Tokyo 112-0012, Japan
tsuda@gssm.otuka.tsukuba.ac.jp

**Abstract.** For the business communication by e-mail with cellular phones, we think there is a serious problem. We sometimes misunderstand the intention of a sender and get into trouble as a result. Such missing each other is caused by a human egocentrism. Therefore, this egocentrism becomes a prevention factor of the good communication on e-Business. To handle this problem, this paper proposes a new model of the egocentrism for e-Business on the Internet based on DSM method which is the standard evaluation index of mental diseases in the psychiatry. It also describes a technique to predict egocentrism using N-gram model. Finally, we examined an evaluation experiment and got results that the precision was 100% and the recall was 34%. From these results, we conclude that this egocentrism presumption method is useful way to reduce misunderstanding the intention of messages on the Internet for e-Business.

**Keywords:** Egocentrism, N-gram, Presumption of egocentrism.

## 1 Introduction

In recent years, e-mail and Web are generally used in e-Business communication by explosive spread of the Internet. In this situation, there are many examples which the intention between senders and receivers is not understood well. For instance, when we use e-mail, we sometimes misunderstand the intention of the sender and get into trouble. Kruger et. al. had reported that it was only 84% which receivers understood the intention of senders [1] for the communication with e-mail. They also showed that such missing was caused by human egocentrism in many cases and confirmed it by some experiments. This egocentrism is mental diseases which cannot throw away own original viewpoint and their concept, and it had been proposed by Piaget [5] who is a child psychologist. Moreover, the egocentrism strongly appears in the days of an infant and it is an emotional character of human which weakens skills in communications with others as they grow up. Such egocentrism is a disincentive factor of a good communication on the Internet besides e-mails. And we experience this in daily life. Therefore, tools and techniques to reduce missing each other of such intention are

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 1002–1009, 2008.

needed. Furthermore, according to the research by Spears et. al., they showed it depends on two following causes to decide whether people go along with another person on the Internet. One is whether a sender is anonymity, another is whether social identity is remarkable at that time or personal identity is remarkable [10]. These properties are strongly related with the egocentrism and we can detect an attitude of the speaker by estimating egocentrism.

On the other hand, Matsumura et. al. distinguished the patterns of the questions behind text by extracting the case flames of verbs [2] for studies to estimate the intention of text messages. Matsumura's technique specifies objects and actions of the intention by superficial language patterns. Therefore, they didn't presume emotional intention such as the egocentrism. In addition, for studies to use egocentrism, Imamura et. al. implemented a humanoid agent with a transactional pattern analysis [4]. Some states called ego states in the agent which expresses the egocentrism are defined in these transactional pattern analysis. They change agent's voice and face expressions with shifting this ego state to an appropriate state by stimulation from users.

In this paper, we focus to the egocentrism in the e-Business communication on the Internet and propose a presumption method of the egocentrism from Japanese text. Concretely, we made a corpus and defined a model for the egocentrism of the text sentences on the Internet. Next, we proposed the method to presume egocentrism by building the presumption rules based on this model and N-gram [7]. Furthermore, we report the evaluation experiment used these rules.

## 2 Building the Rules to Presume the Egocentrism

### 2.1 Building the Corpus for Egocentrism Presumption

In this research, we built a corpus to acquire a tendency of expressions of the egocentrism appeared in text on the Internet. We collected text data in Japanese bulletin board sites which were able to access by cellular phones personally used, because the egocentrism was strongly reflected by individual emotions. Furthermore, we used the bulletin board that served anonymity according to the experimental research that the egocentrism appeared easily in the community on the Internet by Joinson [9]. There were 7,159 sentences in total and 408 sentences (5.7%) that had linguistic expressions specified the egocentrism in these collected data. The judgment of the egocentrism was confirmed by a man and a woman using the egocentrism model described in the following clause. Table 1 shows the part of the collected sentences that include the egocentrism.

### 2.2 Egocentrism Model on the Internet

In this research, we adopted Uchiyama's classification of the egocentrism to develop an egocentrism model on the Internet [3]. This is a classification based on DSM (The Diagnostic and Statistical Manual of Mental Disorders) [6] which is a standard evaluation index of mental diseases in psychiatry. This classification also was evaluated by social psychology experiment of free description questionnaires. For this classification, we expanded the classification items peculiar to the Internet by using

**Table 1.** The example of the egocentrism sentences in our corpus

| Examples of sentences | Description for the egocentrism |
|---|---|
| They are sure to be carrying out the questionnaire which tells the electric wave condition. | It forces own opinion against the other. |
| It is the most certain to confirm it by the telephone or going to the shop. | It forces own opinion against the other. |
| I like black one but hate showing fingerprint. | It complains one-sidedly. |
| So it is impossible with the cellular phone. | It conclude by own judgment. |
| Please understand that it cannot go back and change the request. | It says own reason one-sidedly. |

**Table 2.** The classification of the egocentrism on the Internet

| Classification | Subdivision | Examples of Sentences |
|---|---|---|
| Priority to One's Convenience | Defiant Attitude | Since I am a beginner for using a cellular phone, it is incomprehensible. |
| | Restriction | But it is classic… |
| Self Validity | Conclusion | It is sure to be able to do with all models. |
| | Imposition | You had better send it on repair before the situation turned worse. |
| | Excuse | I was embarrassed because I didn't know how much it was. |
| Belonging to Self-Profit | Demand | I want a device of the feel of a material such as mat because I hate being outstanding. |
| Lack of Empathy | Detachment | Because it has a limit and we cannot play it, I beg you to understand my position. |
| | Irony | It is not a question which you cannot answer without its intention. |
| Attack to Others | Dissatisfaction | I am dissatisfied with the durability of the battery. |
| | Contempt | The antenna of the other company is only useful. |
| Inference | Inference | There might be not different from the waited amount of money ether. |

the corpus that we collected in the preceding clause. The added classifications are "Attack the others" and "Inference". Furthermore, we defined the finer classification items which suited text in the corpus and enabled higher-precision presumption. These subdivisions were created by checking data of the corpus by two or more persons, and we ended up defining 11 items. Table 2 shows these classifications after the expansion.

### 2.3 Building the Egocentrism Presumption Rules with N-Gram

We built the rules to presume the egocentrism on the Internet using collected corpus and the expanded egocentrism model in the preceding clause. We used N-gram method in this study. N-gram is a model Shanon proposed and means the language model that investigates how much frequency N piece of the character string or the combination of the word appears in certain character strings. It is assumed that the occurrence probability of character strings and words depend on the last one in N-gram model as a precondition. Therefore, it is used well in a field of a probability and statistical natural language processing. In this research, we extracted the co-occurrence related character string which appeared the egocentrism at high frequency in the text data which was collected from Web.

At first, we performed a morphological analysis to all sentences which appeared the egocentrism. We used a Japanese morpheme analysis software called ChaSen [8]. This software is most generally used for the research of Japanese natural language processing. Next, we extracted some morpheme rows as the arbitrary N-grams which appeared uniquely and commonly in more than constant number for them. The extraction and

**Table 3.** The egocentrism presumption rules

| Morpheme Rows (In Japanese) *1 | Presumed Egocentrism Classification | | P/N *2 | Strength of the Intention *3 |
|---|---|---|---|---|
| Tadashi (beginning of a sentence) | Priority to One's Convenience | Restriction | N | M |
| Shika/nai/desu/ne | Self Validity | Conclusion | N | M |
| Hazu/desu | | Conclusion | P | M |
| Ga/ii/desu/yo | | Imposition | P | S |
| Beki/[Auxiliary Verb]/yo | | Imposition | P | S |
| Te/shimai/mashi/ta | | Excuse | N | W |
| Te/suimasen | | Excuse | N | S |
| [Particle – Conjunctive particle]/Hoshii/de/sune | Belonging to Self-Profit | Demand | P | W |
| Ashikarazu (end of a sentence) | Lack of Empathy | Detachment | N | M |
| Muri (end of a sentence) | | Detachment | N | S |
| No/ga/zannen | Attack to Others | Dissatisfaction | N | S |
| No/ga/iya | | Dissatisfaction | N | S |
| Shire/mase/n/ne | Inference | Inference | N | W |
| To/omowa/re/masu | | Inference | P | W |
| … | … | | … | … |

*1 "/" describes a separator of morphemes.

*2 "P" is "Positive" and "N" is "Negative".

*3 "S" is "Strong", "M" is "Moderate" and "W" is "Weak".

construction of N-gram was performed from a direction of the begging of the sentence to the end and the opposite direction. We matched these extracted N-grams to expand egocentrism classification items and classified them. Finally, we got 39 rules.

Next, we extracted the P/N (Positive or Negative) classification and the intention strength classification from our corpus for each classification of the egocentrism. We decided whether independence or heteronomy of the intention strength. Each sentence expresses the egocentrism in our corpus often has a positive or a negative meaning. Those sentences also express the strength of the intention of the human who wrote it, and we are able to analyze with high accuracy by making rules from that information. Table 3 shows the part of rules to presume the egocentrism provided in this way. For the intention strength classification in these classification, we defined "S (Strong)" if it is clearly specified positive intention likes or dislikes / right or wrong, "M (Moderate)" if it is specified heteronymous, and "W (Weak)" if it is specified both good and bad.

## 3 The Procedure of the Egocentrism Presumption

The procedures for presuming the egocentrism of a text on Web by using the presumption rules shown in the preceding clause are as follows. In this technique, we don't handle the entire Web page, but presume each sentence on the Web.

[Step 1] One sentence is input from the Web.
[Step 2] The morphological analysis is executed with this sentence.
[Step 3] The appropriate rule is obtained from these morpheme row. It searches from both direction of the beginning and the end of the sentence.
[Step 4] The egocentrism classification is presumed according to the obtained rule and output.

## 4 Evaluation Experiment

We collected 512 sentences from bulletin board sites which were accessible from cellar phones to evaluate this presumption method. These data are separated from the corpus data used to construct the rules. Sentences including the egocentrism that could apply the rule in these data were 94 sentences (17.3%). We applied this method to these 94 sentences and confirmed whether the sentences expressed the egocentrism. As a result, the precision was 100% and the recall was 34% as shown in the equation (1) and (2). Furthermore, F-measure value which is a harmonic mean of the precision and the recall is 50.7% as shown in the equation (3). Table 4 shows the part of obtained presumption result. We realize that the recall in this result is low, because there are few rules that can apply. For this problem, we can deal with increasing the corpus data and the rules for this problem.

$$P_{recision} = \frac{R(Number\_of\_correct\_answers)}{N(Number\_of\_sentences\_for\_applicable\_rules)} = 100\% \tag{1}$$

$$R_{ecall} = \frac{R(Number\_of\_correct\_answers)}{C(Number\_of\_sentences\_having\_the\_egocentrism\_in\_all\_data)} = 34.0\% \quad (2)$$

$$F_{-measure} = \frac{2 \bullet P_{resiion} \bullet R_{ecall}}{P_{resiion} + R_{ecall}} = 50.7\% \quad (3)$$

**Table 4.** The result of the evaluation experiment

| Input Sentence | Applied Rule<br>(In Japanese) | Presumed Result |
|---|---|---|
| I don't know because I don't use it other than that. m(_)m | Mase/n | Self Validity (Excuse)<br>Negative and Strong |
| UTATOMO takes communication fee, but the battery is consumed only if it plays the music. (*^▽^*) | Dake/desu | Self Validity (Conclusion)<br>Positive and Weak |
| "A tragedy of the radio star" may be also good. | Mo/ii/desu/yo | Self Validity (Conclusion)<br>Negative and Strong |
| I am disagreeable if other functions are lost. | No/ga/iya | Attack to Others (Dissatisfaction)<br>Negative and Strong |
| I want a hard lens at least. | [Particle – Conjunctive particle] /Hoshii/de/sune | Belonging to Self-Profit (Demand)<br>Positive and Weak |
| I have a feeling that DRAPE was W46T… | Ki/ga/shi/masu | Inference (Inference)<br>Positive and Weak |
| … | … | … |

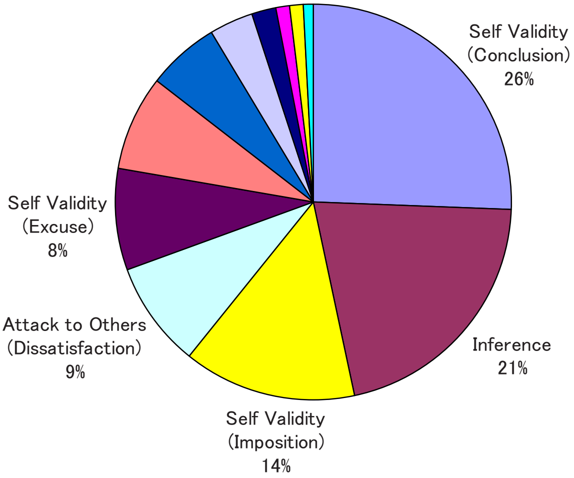


**Fig. 1.** The distribution for the appearance of the egocentrism classifications

In addition, we investigated a tendency to classify the egocentrism in our corpus. As a result, “Self Validity (Conclusion)” was most and it accounted for 26% as shown Figure1. Furthermore, “Self Validity” such as “Conclusion”, “Imposition” and “Excuse” was included within high-ranking 5th place, and it accounted for total 48% that meant a half of all data. On the other hand, “Priority to One’s Convenience” such as “Defiant Attitude” and “Restriction” got the lowest rank. From these phenomena, we realized that the expression of a sentence which had the egocentrism had a lot of “Self Validity” that insisted on own validity. Oppositely, “Priority to One’s Convenience” that insists one-sidedly without considering the circumstance of the partner is few.

## 5 Conclusion

In this paper, we have focused the egocentrism which was one of the obstruction factors of communication on the e-Business. We also proposed its presumption method from real world data. At first we expanded the peculiar to the Internet based on DSM that was the standard evaluation index of the mental diseases in the psychiatry. Then, we defined the ideal classification items for the egocentrism on the Internet. Next, we made the corpus and built the presumption rules which this egocentrism classification corresponded to N-grams of the morphemes. According to the evaluation experiment, although it remains a problem of low recall, we are able to get a performance of 100% for the precision. The misunderstanding of the intention may cause a big damage in the e-Business. Therefore, it is important to understand the egocentrism of the negotiating business partners with such technique for a smooth business talk.

In the future, we have a plan to increase the amount of corpus data and enrich the presumption rules in order to improve the recall. Furthermore, we will examine to expand coverage to the strength analysis of the egocentrism for the document unit such as the Web page. This will be achieved by using the P/N classification and the intention strength classification concerning the egocentrism obtained for this research.

Moreover, when the rules were built from the corpus, we extracted the classification expresses the egocentrism as follows. These are “It is not a yesno answer to a yesno question” for “Priority to One’s Convenience” and “It questioned oppositely though the answer of the question is expected” for “Utilization of Other Person”. These classification needs to analyze the dialogue sentences and we will examine to apply a management method of question states by a stack based structure.

## References

1. Kruger, J., et al.: Egocentrism Over E-Mail: Can we Communicate as Well as We Think? Journal of Personality and Social Psychology 89(6), 925–936 (2005)
2. Matsumura, M., et al.: Extraction of Questions Behind Messages. Transactions of the Japanese Society for Artificial Intelligence 22(1) (2007)
3. Uchiyama, H., et al.: A trial of the narcissism standard making. The 44th Annual Meeting of The Japanese Society of Social Psychology (2003)
4. Imamura, K., et al.: Humanoid Software Agent Applied System. Matsushita Electric Works Technical Report, Vol.53, No.3 (2005)

5. Piaget, J.: Piaget's Theory. Carmichael's manual of child psychology 1 (1970)
6. American Psychiatric Association: Diagnostic and Statistical Manual of Mental Disorders DSM-IV-TR Forth Edition (2000)
7. Mishina, K., et al.: An Emotion Similarity Calculation Using N-gram Frequency. IPSJ SIG Technical Reports, NL-180 (2007)
8. Matsumoto, H.: A morpheme analysis system "ChaSen". Information Processing 41(11) (2000)
9. Joinson, A.: Understanding the Psychology of Internet Behavior: Virtual Worlds, Real Lives (2002)
10. Joinson, A.N.: Understanding the Psychology of Internet Behaviour. Palgrave Macmillan, Basingstoke (2003)

# Generating Dual-Directed Recommendation Information from Point-of-Sales Data of a Supermarket

Masakazu Takahashi[1], Toshiyuki Nakao[2], Kazuhiko Tsuda[1], and Takao Terano[2]

[1] Graduate School of Business Sciences, University of Tsukuba
3-29-1 Otsuka, Bunkyo-ku, Tokyo 112-0012, Japan
[2] Dept. Computational Intelligence and Systems Science, Tokyo Institute of Technology
4259 Nagatsuda-Cho, Midori-ku, Yokohama 226-8502, Japan
{masakazu, tsuda}@gssm.otsuka.tsukuba.ac.jp,
{toshi-nakao, terano}@dis.titech.ac.jp

**Abstract.** Even at the supermarket in Japan, it is commonly used the reward card. However, it is used for only sales expansion objectives with the twice or triple points so far. This paper proposes the methods extracting customer preference information and the characteristics of the commodity from the Point of Sales (POS) data with the reward card. One of the challenges in this paper is how to grasp not only the customer preferences but the trends of the preferences. In the conventional methods, customer preference and market information are managed with two-dimensional vectors of customer and preference category axes. In this proposed method, we add time axis to make it three-dimensional vectors in order to figure out the time-series changes. With this preferences extracting algorithm, we have set up the dual-recommendation site at daikoc.net to browse the trend for both items and customers. Furthermore, we have found trend leaders among the customers, that which confirm that there is a possibility to make appropriate recommendations to the other group member based on the transitions of the trend leaders' preferences.

**Keywords:** Recommendation Systems, Dual-Directed Recommendation, Collaborative Filtering System, Customer Preference.

## 1 Introduction

Information systems are regarded as the effective tools to collect preferences of entire market and individual customer. For the purpose of gathering information about both market trends and customer preferences at once, one of the effective methods is to manage customer preference information using two-dimensional vectors with both customer and preference category axes. This method is used in many of Customer Relationship Management (CRM) tools. However, this conventional two-dimensional method is not capable of keeping track of market trends and the transitions of customer preferences that change over time.

## 2 Objectives for Using the Reward Card at Supermarket

Today, it is commonly used the reward card to collect the customer transaction information even at the super market in Japan. One of the reasons that the store operator

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 1010–1017, 2008.

requests the presentation of the Reward Card at the time of check out is the customers' purchase control. To begin with, some of the roles that the store distributes the Reward Card are for the customer's enclosure and the sales increase. Next reason is for the data analysis to make the best use of the promotion and inventory based on the results of the customers. Generally, the sales in the retail store are resoluble as follows; (Sales) = (Number of the customer transaction) x (Unit price of each customer). Moreover, the customer transaction is resoluble as follows; (Unit price of each customer) = (Unit price of each item) x (Number of purchased items). That is, to increase the unit price or purchased number of items improves the customer transaction. Even when it is difficult to raise a unit price and the number of the customer is also steady, the sales could expand if the customer transaction improves. That is, it becomes the strategy for the sales expansion to buy one more items to the customers. In this sense, recommendation information could make use of the sale expansion. That is, it is profitable for the store operator to use the reward card not only as one of the sales promotion tools but as improvement for the number of purchase items. Therefore, we need to know the preferences for the customer to use for the recommendation information.

## 3 Long Tail Business Issues

Long tail business was referred originally by C.Anderson argued from his book [1] that the products that have low sales volume can collectively make up a market share that exceeds the relatively few current bestsellers, if the distribution channel is large enough. Lots of the recommendation systems in many of the commercial Web sites adopt the collaborative filtering systems. The primary goals of traditional techniques of collaborative filtering systems are improving the accuracy of the recommendation. Although they include many items the users have already known (see Fig.1).

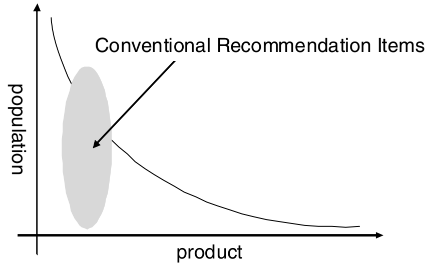


**Fig. 1.** Conventional Recommendation Items among the Long Tail Businesses

In consideration of the accuracy, these recommendations appear adequate. But on the contrary, if we consider users' satisfactions, they are not enough adequate because of the lack of discovery. "Preference" is enumerated in one of the factors to decide one's behavior. Therefore, it is important to model the preference for becoming the variety of basic recommendation information. Among the adjustable recommendation information with user's preference, it is important for recommendation information not only fit user's preference but has both the unexpectedness and the surprise.

## 4 Methods for Customers' Preferences Management

The Customer Preference Database is usually composed of Customer Preference vectors, which are the two-dimensional vectors of customer and category axes. Superposing these Customer Preference vectors implements the three-dimensional vectors with customer, category, and time axes. Each field of Customer Preference vectors contains the score that indicates the strength of preference of a specific customer for a specific category. Tsuda at al [2] formulated the algorithm compresses the information of the time series transition for the items or the users' intentions which have been used in the image data processing. The preference information of unit time is stored as sets of current preference vectors and preference transitions in the past with this algorithm. This three dimensional vector method enables to store information in a small volume and without omission. Making with this algorithm, we have developed a system DIKOC (Dynamic Advisor for Information and Knowledge Oriented Communities: see Fig.2.) with the following functions;

- Recommendation information is generated by the collaborative filtering system based on from both the Point of Sales (POS) data and the reward card data.
- Recommendation method to follow to the information that changes time wise.
- Plot the actual purchase results and recommended items to each customer.

Regarding to the system evaluation experiments has been planned in the shopping area of Hamada city in Shimane prefecture, Japan.

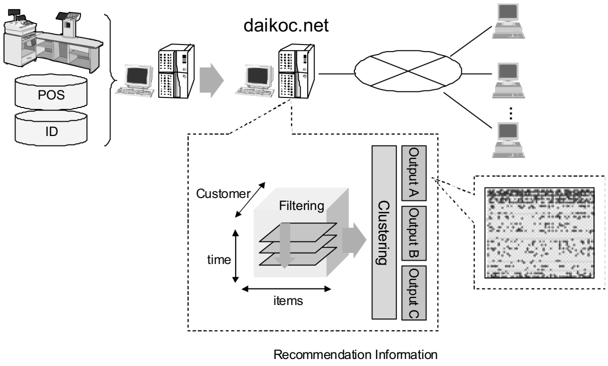
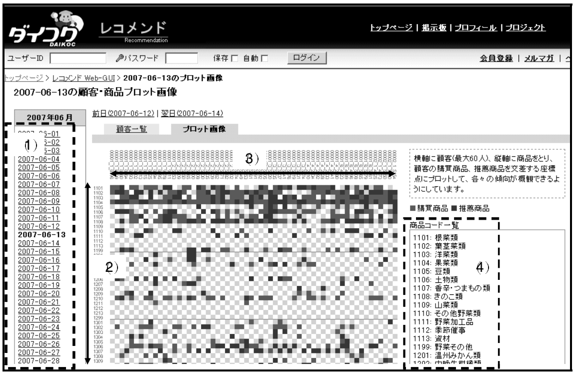


**Fig. 2.** Recommendation Information Generating System at daikoc.net

## 5 Dual-Directed Recommendation Systems

DAIKOC system has the following components with match-making functionalities to bridge gaps among consumer, retailer, and manufacture. Therefore, we referred "Dual-Recommendation" because of providing the recommendation information not only to consumers but also to manufacturers and retailers. The recommendation information system is analyzed from the data of the management system and the Point of Sales (POS) data with the reward card and can offer recommendation information



2007-06-13の顧客・商品プロット画像

横軸に顧客(最大60人)、縦軸に商品をとり、顧客の購買商品、推薦商品を交差する座標点にプロットして、各々の傾向が概観できるようにしています。

商品コード一覧

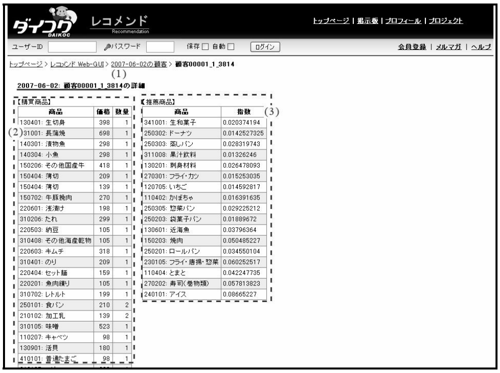
| | |
|---|---|
| 1101: | 根菜類 |
| 1102: | 葉茎菜類 |
| 1103: | 洋菜類 |
| 1104: | 果菜類 |
| 1105: | 豆類 |
| 1106: | 土物類 |
| 1107: | 香辛・つまもの類 |
| 1108: | きのこ類 |
| 1109: | 山菜類 |
| 1110: | その他野菜類 |
| 1111: | 野菜加工品 |
| 1112: | 季節催事 |
| 1113: | 資材 |
| 1199: | 野菜その他 |
| 1201: | 温州みかん類 |

(1) Date axis, (2) Categories axis, (3) Clients axis, (4) Category List

**Fig. 3.** Recommendation Information at daikoc.net

including time series changes to both an individual consumer and the supermarket manager. Fig.3. shows the site image for the recommendation information system. This shows the plot image for both the actual purchase and the recommended items to each customer. The vertical axis show the category for the items and the horizontal axis indicates the customer. From this page, we can figure out the tendency view for the relation both of the purchased items and the recommended items.

Not only indicates the whole trend in this site, this system has the search function for the personal activities that show the display for both the actual purchases and the recommended items of a certain customer. Fig. 4 shows the display for both the actual purchase and the recommended items of a certain customer. From the actual purchase matrix, we can figure out items, prices, and quantities. The recommended items show the index for the strength of the recommendation, respectively. From this recommendation information generates the new knowledge for customer, retailer and manufacture. For example, both retailer and manufacture will make use for the inventory control. However, this information only indicates to whom and to which. But it is one of


ダイコク
DAIKOC
レコメンド
Recommendation
トップページ | 掲示板 | プロフィール | プロジェクト
ユーザーID
パスワード
保存 □ 自動 □
ログイン
会員登録 | メルマガ | ヘルプ
トップページ > レコメンド Web-GUI > 2007-06-02の顧客 > 顧客00001_1_3814
(1)
(2)
(3)

2007-06-02: 顧客00001_1_3814の詳細

【購買商品】

| 商品 | 価格 | 数量 |
|---|---|---|
| 130401: 生切身 | 398 | 1 |
| 131001: 長蒲焼 | 698 | 1 |
| 140301: 漬物魚 | 298 | 1 |
| 140304: 小魚 | 298 | 1 |
| 150206: その他国産牛 | 418 | 1 |
| 150404: 薄切 | 209 | 1 |
| 150404: 薄切 | 139 | 1 |
| 150702: 牛豚挽肉 | 270 | 1 |
| 220601: 浅漬け | 198 | 1 |
| 310206: たれ | 299 | 1 |
| 220503: 納豆 | 105 | 1 |
| 310408: その他海産乾物 | 105 | 1 |
| 220603: キムチ | 318 | 1 |
| 310401: のり | 209 | 1 |
| 220404: セット麺 | 159 | 1 |
| 220201: 魚肉練り | 105 | 1 |
| 310702: レトルト | 199 | 1 |
| 250101: 食パン | 210 | 2 |
| 210102: 加工乳 | 139 | 2 |
| 310105: 味噌 | 523 | 1 |
| 110207: キャベツ | 98 | 1 |
| 130901: 活貝 | 180 | 1 |
| 410101: 普通たまご | 98 | 1 |

【推薦商品】

| 商品 | 指数 |
|---|---|
| 341001: 生和菓子 | 0.020374194 |
| 250302: ドーナツ | 0.0142527325 |
| 250303: 蒸しパン | 0.028319743 |
| 311008: 果汁飲料 | 0.01326246 |
| 130201: 刺身材料 | 0.026478093 |
| 270301: フライ・カツ | 0.015253035 |
| 120705: いちご | 0.014592817 |
| 110402: かぼちゃ | 0.016391635 |
| 250305: 惣菜パン | 0.029225212 |
| 250203: 袋菓子パン | 0.01889672 |
| 130601: 近海魚 | 0.03796364 |
| 150203: 焼肉 | 0.050485227 |
| 250201: ロールパン | 0.034550104 |
| 230105: フライ・唐揚・惣菜 | 0.060252517 |
| 110404: とまと | 0.042247735 |
| 270202: 寿司(巻物類) | 0.057813823 |
| 240101: アイス | 0.08665227 |

(1) Date, Customer ID, (2) Purchased items, (3) Recommended items

**Fig. 4.** Detailed Recommended Information

the important things to know the timing for the recommendation even among the retail business field. Therefore, we need to know the timing for the recommendation. Moreover, in the case of the recommendation for the commodity items, we should take it consideration both to the dealing axis and sequential axis. Because, something of heavy or bulky stuff such as Rice, Food Oil, Mayonnaise, and Wine that may purchase repeatedly and differs from such as book recommendation. Because the commodity items differ order of the price from the book and the book doesn't buy the same one if we buy it once, but the commodity items are bought repeatedly depending on the preference. Moreover, the commodity recommendation occasionally is recommended some different items or same items that were bought before.

Table1 shows the recommendation results from the evaluation data that gathered in a supermarket. From this table, our engine forecasted the accuracy with 6.8%. In this experiment, we have used the Taste [4] for the original recommendation engine. Taste is an open source flexible collaborative filtering engine for Java. This engine takes users' preferences for items and returns estimated preferences. We have improved the engine for the accuracy improvement of the recommendation probability. Our system recommends the items with high collocation from the past transaction data like amazon.com. Typical index for collocation are as follows; (a) co-occurrence frequency, (b) Jaccard coefficient, (c) Simpson coefficient, (d) cosign distance. Among them, this time we take cosign distance.

**Table 1.** Recommendation Results

| Method | Similarity Method | # of items | Forecasted items | Accuracy |
|---|---|---|---|---|
| User-based (Taste) | Correlation | 3,190 | 5 | 0.2% |
| User-based (Original) | Cosign | 3,190 | 219 | 6.8 % |

## 6 Extracting the Trend Leaders' Activities

The timing which item when to recommend to the customers can grasp by searching for the purchase items those whose activities are in the state of the arts. Hereafter, we define the trend leaders as follows; The Group those who have already purchased the item that will be in fashion long before. Namely, our objective is for extracting the items that maximizing the potential customers from the trend leaders' activities. That is, we should look out the customer who has bought the item that is currently in the high Purchase Index (PI) before the score the rises. The Purchase Index (PI) that stands for the number of the sales items per 1,000 customers. Define (Purchase Index) = (Number of the sales items) / (Number of the check out customer) x 1000. This index is frequently used among the retail business. At first, we select the items that have the high Purchase Index (PI) in the present time. Among them, we extract some items those are increasing in Purchase Index (PI) by time series and then sort out the customers that purchase those items. From the activities of the trend leaders that when those trend leaders have purchased the items that raised Purchased Index (PI), we can

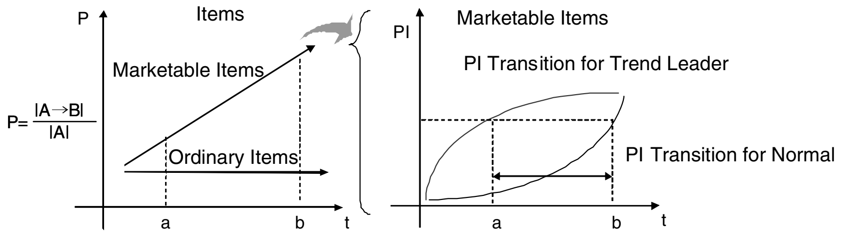


**Fig. 5.** Method for Finding out the Trend Leader

grasp the cutting edge of the trend (see Fig. 4.). Namely, this extracting steps result in discovering the combination of items that maximize the number of the potential customers.

Assuming person *l* bought the item *i* at time (*t-j*) that the item that has started selling at time (*t-k*), let $p_{i\quad l}$ define the purchase amount of the item $p_i$ for person *l.* Given that $p_i^{t-j} < p_i^t$ , then $p_{i\ l}^{t-j}$ is defined at the time (*t-j*) of the purchase amount of item $p_i$ for person *l.*

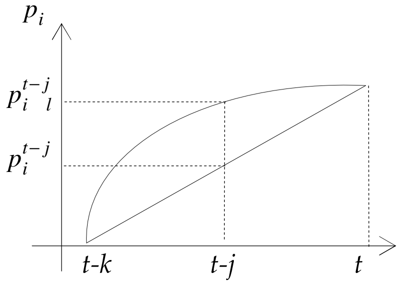
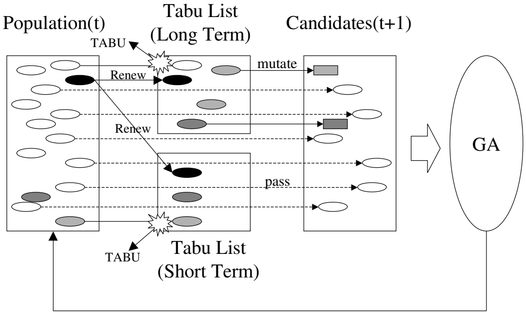


**Fig. 6.** Time Series Transition for the Inclination of Sales

The item monetary weight at the initial stage is given $j_i/k_i$ , and the inclination of sales for item $p_i$ follows $p_i^t/p_i^{t-j}$ . Hence, the following formula $\sum_i \frac{p_i^t}{p_i^{t-j}} \frac{j_i}{k_i} p_{i\ l}^{t-j}$ is given the answer for the customers who have bought the recommendable items long before (see Fig.7). With this formula, we extract the customers who have bought future recommendable items. The following examine is researching the items that the trend leaders have been bought and extracting the items and the conditions of becoming the trend leaders. There are about 20,000 items handled in the store, For example, even though the two items association rules have to calculate at least 400,000,000 combinations. Moreover, most of the stores operate all the year and about 10,000 customers are held by at least 1,000 sq. meter large type store. Therefore, it will take time to solve in real time calculation, so that we make use of GA calculations. This time, we examine TABU-GA [3] to extracting the trend leaders' consumption activities. Multipronged optimal solutions can be led with this GA algorithm, even though with the conventional GA algorithm can be led one optimal answer. The main idea of the algorithm is that, (a) in each generation, one best individual generated by GA operation is stored into the tabu-lists to inhibit it from selecting specified times, and (b) solution candidates found in the previous generations will become tabus, and thus, the other candidates are explored in order to get better and divergent solutions. Fig. 6 shows the outline for the extracting steps.



1. Set $H$ Empty, $H$ which is a historical memory. Select $x^{now} \in X$ as an initial solution.
2. Choose $selection\_N(x^{now}) \subset N(H, x^{now})$, where $N(H, x^{now})$ is a set in $x \in X$ except in the neighborhoods of $H$.
3. Select $x^{next} = \max(c(H, x^{now}))$, $x^{next} \in selection\_N(x^{now})$, where $c(H, x^{now})$ is an objective function is a mapping of a set in $x \in X$ except in the neighborhoods of $H$.
4. Run GA.
5. If a condition of ending is true then end
6. Exchange $x^{best}$ for $x^{old}$ in $H$. Return to 2.

**Fig. 7.** Algorithm for Tabu-GA

With these algorithms, we will extracting the trend leaders and find the items that are recommendable that will be scored in high Purchase Index (PI) in the future.

## 7 Conclusion

In this paper, a basic research project in relation to the Point-of-Sales (POS) transaction data analysis with the dual-directed recommendation was described. Finding out the formula for extracting the trend leaders among the customers, that which confirm that there is a possibility to make appropriate recommendations to the other group member based on the transitions of the trend leaders' preferences. Though the Amazon.com is famous as the case with the recommendation, according to the patent document, the items marked high score were purchased lot [5]. As the factors of high accuracy recommendation with the Amazon.com, it is pointed out following issues; the evaluation data after the purchase from the users, collecting five stage evaluations from the user reviews, and so on. It is important to generate the evaluation index to progress the recommendation accuracy.

This research was partly with the support of The Ministry of Economy, Trade and Industry (METI), the Council on Improvement of the Service Industry's Productivity (in the field of service engineering research and development, and the field of service engineering application demonstration). We express appreciation to those involved.

## References

1. Anderson, C.: The Long Tail: Why the Future of Business Is Selling Less of More, Hyperion (2006)
2. Tsuda, K., Hirano, T., Takahashi, M., Terano, T.: 3-D Knowledge Structures for Customer Preference Transition. In: IEEE Int. Conf. on Systems, Man and Cybernetics (2002)
3. Takahashi, M., Kurahashi, S.: Tabu Search Algorithms for Multimodal and Multi-Objective Function Optimizations. IJCSNS 7(10), 257–264 (2007)
4. Taste, http://taste.sourceforge.net
5. Linden, G., Smith, B., York, J.: Amazon. Com Recommendations; Item-to-Item Collaborative Filtering. IEEE Internet Computing, pp. 73–80 (January-February, 2003)
6. Hijikata, Y., et al.: Special issue- Capture and Make the best use of the Users' Favours - Front Line of the Preferences Extraction Technologies. IPSJ Magazine 48(9) (2007) (in Japanese)
7. Pei, M., Taniguchi, S., Hara, T., Nishio, S.: Association Rule Mining Considering Repetition in Purchase to Discover Important Association Rules and Loyal Customers. Journal of IPSJ 47(12), 3352–3364 (2006)
8. Kessoku, M., Takahashi, M., Tsuda, K.: A Method of Customer Intention Management for a My Page System. In: 8th International Conference on Knowledge-Based Intelligent information Engineering Systems, pp. 523–529 (2004)
9. Adomavicius, G., Tuzhilin, A.: Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. IEEE Trans. on Knowledge and Data Engineering 17(6), 734–749 (2005)
10. Schafer, J.B., Konstan, J.A., Riedl, J.: E-Commerce Recommendation Applications. Data Mining and Knowledge Discovery 5, 115–153 (2001)
11. Burke, R.: Hybrid Recommender Systems: Survey and Experiments. User Modeling and User-Adapted Interaction 12, 331–370 (2002)
12. Herlocker, J., Konstan, J., Terveen, L., Riedl, J.: Evaluating Collaborative Filtering Recommender Systems. ACM Transactions on Information Systems 22(1), 5–53 (2004)
13. Maglio, P.P., Srinivasan, S., Kreulen, J.T., Spohrer, J.: Service Systems, Service Scientists, SSME, and Innovation. Communications of the ACM 49(7), 81–85 (2006)
14. Kovacs, T.: Strength or Accuracy: Credit Assignment in Learning Classifier Systems (Distinguished Dissertations). Springer, Heidelberg (2004)

# Extraction of the Project Risk Knowledge on the Basis of a Project Plan

Yasunobu Kino, Kazuhiko Tsuda, and Tadashi Tsukahara

Graduate School of Business Sciences, University of Tsukuba
3-29-1 Otsuka, Bunkyo-ku, Tokyo 112-0012, Japan
kino@mbaib.gsbs.tsukuba.ac.jp,
tsuda@gssm.otsuka.tsukuba.ac.jp,
t-tsuka@poem.ocn.ne.jp

**Abstract.** When identifying project risks, a checklist created from the former projects tends to be used. However, each project has its uniqueness. If we would like to make a check list comprehensive and available for every project, it would end up containing large amount of items to be checked. Also that type of checklist could prevent a project manager from considering its essential risks in the project and make the risk identification process routine. As a result, it could decrease the effectiveness of risk management, especially when it's done by an entry-level project manager. In this research, methods for creating a useful checklist to reduce these problems are discussed, by using knowledge based risk identification support system which extracts a few yet important check items depending on the characteristics of a target project.

**Keywords:** Project Management, Risk Management, Assumption Analysis, Project Planning.

## 1 Introduction

To manage software development appropriately, it is important to recognize and manage existing risks in the project in order to keep problems from occurring and to minimize the influence even when they actually happen. With those reasons, risk management has been recognized as one of the important knowledge areas in the project management field. Therefore, risk management is essential to the success of projects, and many case studies on project risk management have been reported [1],[2],[3]. It is necessary for risk management to identify hidden risks in the project appropriately in advance. There are three common methods for risks identification; “Checklist”, “Interview to Experts”, and “Brainstorming”. They are very effective to identify risks, but each of them also has its limitations. For instance, checklists tent to be described in abstract forms so that they can be applied to any type of project. As a result, items in the checklist can easily be ambiguous. The methods “Interview to experts” and “Brainstorming” depend significantly on the skills and experiences of a project manager. It is very likely that we just don't realize risks because we have

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 1018–1025, 2008.

never experienced similar type of projects. Or we might have a big difficulty in identifying risks appropriately in the project due to the lack of experience.

In this research, we discuss about knowledge based risk identification support system as a tool to help resolving these issues. This knowledge based risk identification support system provides a capability to extract a list of important potential risks. A list can be extracted by entering distinctive attributes of the target project. The list looks similar to the traditional checklist. But it is different for the point that the list extracted from knowledge based risk identification support system only contains important and adequate number of risks based on certain target project characteristics.

## 2 Confirmation for Project Risks

In this chapter, we examine to confirm the variety of risks extracting from the list of issues and concerns gathered from the actual project.

## 3 Extracted Risks from the Issues and Concerns Table

Prior to this research, we first obtained a list of issues and concerns described by project members of an actual software development project. Then we extracted risks from the list to see what kinds of risks exist in an actual project. Other than potential risks, the list also contains memos and the description of problems that already happened. Since the purpose of this research is to know what kinds of risks exist in a project, we considered following points when we check the issues and concerns in the list.

Points considered are as follows;

- Repeated issues and concerns were counted once.
- Technical issues and concerns found in tests and trial procedures were eliminated except typical one.
- Notes and memos were eliminated as they are not recognized as risks.

From the documents from the issues and concerns list, there are 25 risks were extracted as a result of the risk extraction. Table 1 shows the risks. The issues and concerns list are described only from engineers' perspectives. To make the list of risks more comprehensive, we considered some additional perspectives such as occupational health and safety, human behavior and crime control. After that, total 38 risks were extracted as a result.

## 4 Analyzes of Risk Factors

Next, we reviewed these risks and analyzed their causes by using the affinity diagram. Figure 1 shows the analysis procedure.

Thinking about the causes of risks in Figure 1, we found that the causes of risks were categorized into two causes; one is the cause rooted in the project components, and another is the cause rooted in fundamental causes of risks or mechanism of risk occurrence. Moreover, we found that each risk description always involves those two

**Table 1.** Extracted risks from issues and concerns table

| No | Risks |
|---|---|
| 1 | There might be some possibilities of the delay deliveries for software on test version. |
| 2 | There might be some possibilities to switch the software easily for the parameters and the programs, because of the information delivery that used on the test version. |
| 3 | There might be some possibilities to fail into the connection on testing the network connection, |
| 4 | Unknown for the workload that changes for the I/O definition. |
| 5 | Not decided the revision and confirmation for the performance parameter. |
| 6 | Not decided the methods for version up of the security management software. |
| 7 | Not both guaranteed and recorded operating results for third parties software |
| 8 | There might be some possibilities for the disk spaces shortage. |
| 9 | There might be some possibilities for increase the workload for the unknown new additional warning alerts. |
| 10 | There might be some possibilities for the software performance stabilities after the cut-over. |
| 11 | There might be some possibilities for the delay the software developments for delivering the parts lists or the APL lists to the cooperate companies. |
| 12 | There might be some possibilities for miscoding 4 to ■ at the letter recognition |
| 13 | There might be some possibilities for failure the work flow analysis with the conventional middle ware functions. |
| 14 | There might be some possibilities for recognition error with space code on the documents |
| 15 | There might be some possibilities for failure the recognition with "\" letter on the documents. |
| 16 | Not defined the name of the bank, the branch and the subject. |
| 17 | Not operated the sample programs that have already obtained. |
| 18 | There need to add some functions for obtaining the coordinates without analysis in the documents analysis on the attributes definition. |
| 19 | There need to add some interface functions for the document analysis ratio in the application side control. |
| 20 | Add the documents rote functions on the screen in the application control side |
| 21 | There might be some possibilities for the delay because of the information for updating the image database attributes delayed delivery. |
| 22 | There might be some possibilities for the delay because of the worse software test performance. |
| 23 | There might be some possibilities for confusion because of the vague definition with the attributes changes both the February version and the March version. |
| 24 | There might be some possibilities for the lost communications because of not fixed the specification for OCX in the case of the software revision. |
| 25 | There might be some possibilities for failure for the use of the document Image data that put them before in the case of the software version up. |

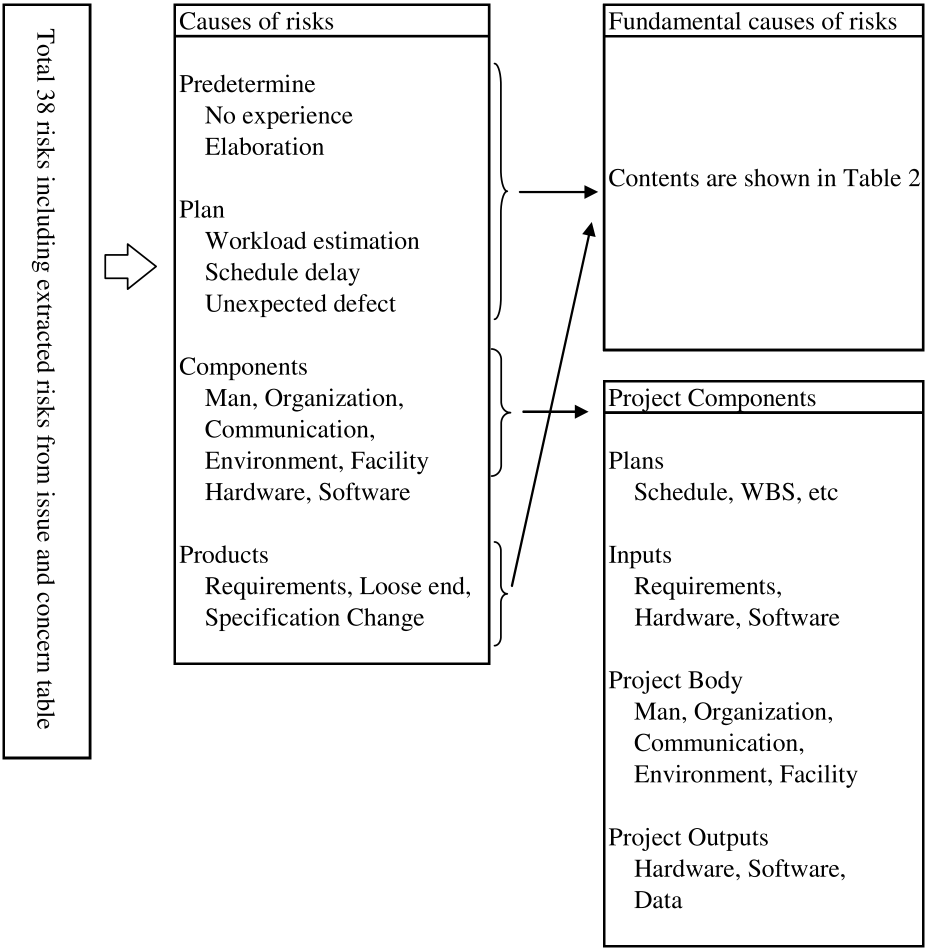


**Fig. 1.** Thinking processes about causes of risks

elements, project components and fundamental causes of risks. For Example, in case of "Potential shortage of disk drive space", "space" or "disk drive space" is corresponding to "project components". "Potential shortage" is corresponding to "fundamental causes of risks". In other words, it can be addressed as "More space might be needed than planed". Table 2 shows the analysis of causes of risks by using affinity diagram [4]. As b) shows, the situation of "More space might be needed than planed" could be caused by the gap between the initial plan and the actual result.

Every project has a plan. If the project plan was perfect and project was assured to be executed just as the plan, there would be no risks except failures, errors and external factor. But there is no project that is free from assumptions and assumptions always contain uncertainty. Uncertainty could cause gaps between the plan and the actual result. This is one of the major reasons why every project has risks.

**Table 2.** The taxonomy of fundamental causes of risks

a) No or few experience  
b) Risk identified in project operation  
b-1) Unexpected and hidden tasks and problems.  
b-2) Gaps between the initial plan and the actual result.  
c) Constraints, inconsistencies, mistakes  
Machines, costs, timeframe, quality, specifications  
d) Failure, error (Internal factors)  
Machine trouble, Human error, internal crime  
e) External factor  
Country risk, monetary exchange, bankruptcy

## 5 Risk Definition

As discussed in the previous section, we found that each cause of risks can be categorized in either the cause related to project components or fundamental causes of risks. This fact was found from analysis on the causes of risks by using the affinity diagram. There are some differences in definitions of risk among research areas. For example, AS/NZS 4360[5] defines risks as “the chance of something happening that will have an impact upon objectives. It is measured in terms of consequences and likelihood”, and ISO/IEC Guide 73[6] defines risks as “Risk can be defined as the combination of the probability of an event and its consequences”.

These definitions tell us that probabilities and consequences are important factors when we describe certain risks. On the other hand, the risk descriptions used in this research have no description on its probabilities and consequences. But risk analysis and evaluation are normally performed after risk identification process, and probabilities and consequences are evaluated in those processes. Therefore probabilities and consequence don’t have to be described in the risk identification process.

## 6 Risk Generation and Extraction

Causes of potential project risks can be categorized either in “project components” or in “fundamental causes of risks”. It means that these potential risks can automatically be generated by these two elements. Figure 2 shows the mechanism of automatic risk generation. In this matrix, the horizontal axis shows “project components” and the vertical axis shows “fundamental causes of risks”.

It is difficult to use this mechanism in an actual project yet because too many potential risks can be generated. Risks to be extracted should be limited. In this research, we tried to extract 30 important risks, as we know from experience that it is not so easy to manage too many risks.

The numbers of team members engage in software development projects vary, from a few members to over 1000 members. As a matter of course, project characteristics also vary by project size and a risk could have different meaning depending on

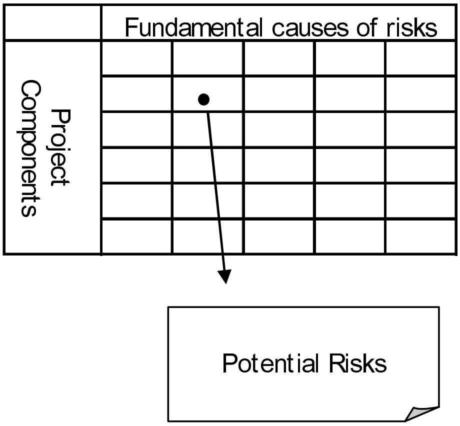


**Fig. 2.** Mechanism of Automatic risk generation

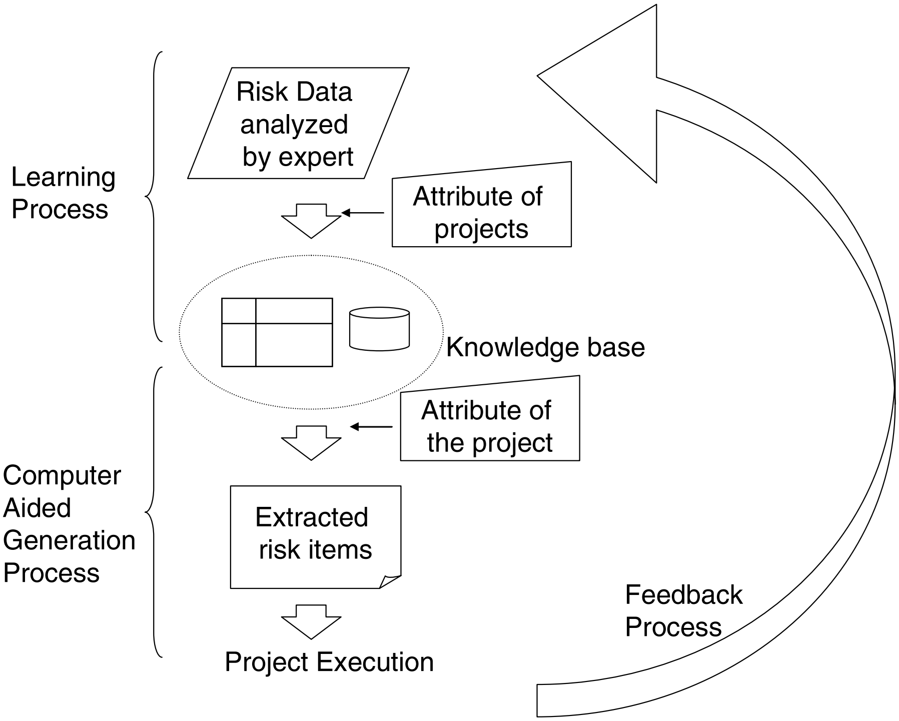


**Fig. 3.** Outline of Knowledge base risk identification support system

project characteristics. Then, it is important to change the meanings of risks depending on project attributes. Figure 3 shows the outline of Knowledge base system.

The first step is the learning process of knowledge base risk identification support system. In this process, project experts extract important risks from the data of previous projects and accumulate them as a knowledge base. In this phase, project attributes are also defined. Next step is the computer aided generation process. In this process, the knowledge base risk identification support system generates important risks by using certain project attributes. And the third process is feed back process. In this process, we compare the generated risks with actual result and feed back the information to the system.

The risk list extracted from the knowledge base system is similar to commonly used checklist. Differences between these two lists are; the system generated list is specialized for a target project, and important risks are listed and numbers of generated risks are limited.

The characteristics of this knowledge base risk identification support system are;

1. Generate important limited number of risks. The first target is 30 risks.
2. Generated risks can be described as “Project components” have “fundamental causes of risks”.
3. Project plan is one of the important project components that could generate risks.
4. Knowledge base can be improved by data accumulation

## 7 Conclusion

Through the process of analysis on risks recognized in projects, we found that the causes of project risks can be expressed as “Project components” and “fundamental causes of risks”. Using this fact, we discussed the mechanism of a knowledgebase that automatically describes potential risks in a project and extracts important risks.

The remaining considerations are how to categorize project attributes and how to weigh the importance of risks. The next step is to develop the knowledge based risk identification support system, accumulate data, and utilize it in actual projects.

## References

[1] Williams, R.C., Walker, J.A., Dorofee, A.J.: Putting Risk Management into Practice. IEEE Software (May/June 1997)

[2] Yokota, T., Kawabata, K., Niino, T., Kawasaki, T.: Development of a Contract Risk Assessment Support System. Journal of the Society of Project Management 7(3), 20–25 (2005)

[3] Muramatsu, M., Okamura, T.: Risk Management in the Projects for Development of an Information system. Journal of the Society of Project Management 9(4), 18–22 (2007)

[4] Kino, Y.: A Proposal of Risk Event Identification Method using Two-dimensional Table composed Project Component and Classified Perils. Journal of the Society of Project Management 3(6), 28–33 (2001)

[5] Standards Australia and Standards New Zealand, Australian/New Zealand Standard 4360:1995 Risk Management (1995)

[6] ISO/IEC, ISO/IEC Guide 73:2002 Risk management – Vocabulary – Guidelines for use in standards (2002)

[7] A National Standard of Canada, CAN/CSA-Q850-1997 Risk Management: Guideline for Decision-Makers (1997)

[8] IEEE Computer Society, IEEE Std 1540-2001:IEEE Standard for Software Life Cycle Processes – Risk Management (2001)

[9] Kirkpatrick, R.J., Walker, J.A., Firth, R.: Software Development Risk Management: An SEI Appraisal. Software Engineering Institute Technical Review 92(CMU/SEI-1992-REV) (1992)

[10] Carr, M.J., Konda, S.L., Ulrich, F.C., Walker, C.F.: Taxonomy Based Risk Identification Technical Report CMU/SEI-93-TR-6, Software Engineering Institute, Carnegie Mellon Univ. (1993)

[11] Software Engineering Institute, Continuous Risk Management Guidebook, Software Engineering Institute, Carnegie Mellon Univ. (1996)

[12] Jones, C.: Assessment and Control of Software Risks. Prentice Hall, Englewood Cliffs (1994)

[13] Meredth, J.R., Mantel Jr., S.J.: Project Management - A Managerial Approach. John Wiley & Sons, Inc., Chichester (1995)

[14] Moynihan, T.: How Experienced Project Managers Assess Risk. IEEE Software, 35–41 (May/June 1997)

[15] Thayer, R.H., Fairley, R.E.: Software Risk Management, Software Engineering Project Management, 2nd edn. IEEE Computer Society Press, Los Alamos (1997)

[16] Boehm, B.W., DeMarco, T.: Software Risk Management. IEEE Software 14(3), 17–19 (1997)

[17] Smith, P.G., Merrit, G.M.: Proactive Risk Management: Controlling Uncertainty in Product Development, Productivity Pr. (2002)

[18] DeMarco, T., Lister, T.: Risk Management during Requirements. IEEE Software 20(5) (2003)

[19] Kado, K., Horiuchi, T., Seki, T.: Application of FMECA to Project Risk Identification Process. Journal of the Society of Project Management 5(2), 19–25 (2003)

[20] Okada, K.: Risk Management in IT Solution Projects - A Study for Bridge between Knowledge and Actual Cases. Journal of the Society of Project Management 9(4), 23–28 (2007)

[21] Yamato, S.: Project Assessment Methodology as One of PMO Functions. Journal of the Society of Project Management 9(4), 41–46 (2007)

# Real-Valued LCS Using UNDX for Technology Extraction

Setsuya Kurahashi

University of Tsukuba, 3-29-1 Otsuka, Bunkyo, Tokyo 112-0012, Japan
kurahashi@gssm.otsuka.tsukuba.ac.jp

**Abstract.** This paper proposes a new method of developing a process response model from continuous time-series data. Chemical and/or biotechnical plants always generate a large amount of time series data. However, since conventional process models are described as a set of control models, it is difficult to explain complicated and active plant behaviors. To uncover complex plant behaviors, the method consists of the following phases: (1) Reciprocal correlation analysis; (2) Process response model; (3) Extraction of a workflow; (4) Extraction of control rules of real-valued data. The main contribution of the research is to establish a method to mine a set of meaningful control rules from a Learning Classifier System using UNDX(Unimodal Normal Distribution Crossover) for real-valued data. The proposed method has been applied to an actual process of a biochemical plant and has shown its validity and effectiveness.

## 1 Introduction

So far, many kinds of automatic control systems have been established in such plants as chemical plants. Operator confirmation and manual procedures are essential for a wide variety of products used in small quantities requiring stringent quality control, such as advanced materials for Liquid Crystal Display (LCD), pharmaceutical products, and so on. The quality control of biochemical plants has also become one of the most important issues in the field of food-safety.

In the past, transfer functions like the delay time function have built up a process model by describing an individual response process. The transfer function is used in analyses of input/output behaviors. It is derived using the Laplace transform in control theory. However, process circumstances might change significantly, according to variations of infused material or operating conditions. Thus, automated acquisition or data mining of processes from actual daily data is desirable to manage these changes. In this research, we propose a heuristic search method for plant operation rules, which could provide guidance on human operators, building up a process response model from a large amount of time series data. The basic principles of the model are 1) to maximize the correlation coefficient among time series data, 2) to apply LCSs with Minimum Description Length (MDL) criteria [1], and 3) to apply Genetic Algorithm with Unimodal Normal Distribution Crossover(UNDX). The paper also describes results from applying the proposed method to actual operation data for a biochemical plant.

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 1026–1033, 2008.

# 2 Research Objective

## 2.1 A Target Plant

We are dealing with a biochemical plant with a distillation tower. In the distillation tower, low-pressure treatment performs constituent separation after the basic ingredient is infused into the tower.

Simply observing normalized data is not enough, and it is almost impossible to read what kind of relationship exists among process data. Operators, who are those who actually operate the plant and learn about the process characteristics empirically, control the whole plant by frequently adjusting control settings or by manual operation.

## 2.2 Problem Description

The purpose of this research is to extract significant information from such time series data that appear to be complicated. The following phases show how to build a model up in order to analyze the process data.

1. Process data acquisition phase
   Various kinds of process data are collected and stored in a database.
2. Normalization phase
   Each piece of process data has a different range.
3. Reciprocal correlation analysis phase
   Two sets of normalized process data are selected and searched for the time difference that indicates the biggest correlation between each process value by gradually shifting the time.
4. Process response model phase
   Like the response model, the phase describes the relationship among the process data from the shifted time and the correlation coefficient.
5. Extraction of a control rule phase
   The process extracts the control rule by executing LCS that handles specified process data as a process response model class.

# 3 Principles of LCS with MDL

## 3.1 MDL Criteria

Operation rules of such plants require simple and clear descriptions in order for operators to recognize the target process conditions. The concept of complexity determines the data and the model describes it [2]. MDL criteria minimize the complexity of the model and data [1,3]. We use it to minimize the complexity of LCS. MDL criteria are shown as follows [4]: Here, $m_1$ and $m_0$ are each occurrence numbers of $y = 1$ and $y = 0$, in data row $y^m = y_1, ..., y_m$ with length $m$. Here, $m = m_1 + m_0$. Also, $c_i$ indicates 0 where each condition of the former part has a wild card #, or 1 in other cases. And $t_i$ indicates the division number of the process data in each condition; $k$ indicates the number of conditions. Then, the description length of data and model are as follows:

$$\begin{aligned} dataLength &= mH(\frac{m_1}{m}) + \frac{1}{2}\log(\frac{m\pi}{2}) + o(1), \\ modelLength &= \sum_{i=1}^{k} c_i(1 + \log t_i), \\ where, H(x) &= -x\log(x) - (1-x)\log(1-x). \end{aligned}$$

### 3.2 Improvement Rate Based MDL Criteria

Although the MDL principle generates a simple and safe model, this does not always means the model is easy to understand. Thus, we apply the improvement rate of association rules in data mining literature [5]. The following formula expresses the improvement rate:

$$improvement = \frac{P(r_i|\mathbf{p})}{P(r_i)},$$

where, $P(r_i)$ expresses the rate that the latter part $r_i$ appears without condition, and $P(r_i|\mathbf{p})$ expresses the rate that the latter part $r_i$ appears with the condition of the former part, $\mathbf{p}$.

The description length in MDL principles is calculated as follows: It is known that probability distribution $P(\cdot)$, on the assembly of data row $y^m = y_1, ..., y_m$ with length $m$, exists. Also, the length $L(y^m)$ of binary code string $\phi(y^m)$ can be expressed as $L(y^m) = -\log P(y^m)$. Expressing the occurrence rate in marketing basket analysis through a logarithm with the same description length as the description length of MDL principles, it is possible that the improvement rate is the differential of the information amount between before-refining and after-refining, with certain conditions. Therefore, we handle the improved information amount and the description length in MDL principles at the same time. So, MDL criteria are expanded, in order to maximize the differential of the description length (model length + data length) obtained for classification by LCS. The following method shows the calculation of the learned classifier weight. Here, $dataLength_f$ and $modelLength_f$ express the initial description length of data and model, $dataLength_l$ and $modelLength_l$ express the final description length of data and model. The weight of classifier is calculated by

$$\begin{aligned} DL_{first} &= dataLength_f + modelLength_f, \\ DL_{last} &= dataLength_l + modelLength_l, \\ Weight &= DL_{first} - DL_{last}. \end{aligned}$$

When applying a rule and the rule is simple, the knowledge, which is unknown when there are no rules, reveals another unpredictable rule [6]. The expansion proposed here allows for detailed evaluation of the simple rule that can reveal a valuable fact with copious amounts of information. Evaluating all classifiers hit in the former part, counting the result of its classifier allows for calculation of the estimated value of a classification error. MDL criteria are used to get the weight, selecting the smallest dataLength in the classification of results. Then the learning classifier system is implemented to maximize the weight as a fitness function.

### 3.3 Learning Classifier System

As a learning classifier system, a modified system based on the original LCS is introduced [7,8,9,10]. This system corresponds to a lot of events, using the learning method it estimates event distribution by random sampling. Each individual consists of a condition part (as *disjunctive normal form*) and of a conclusion. First of all, random generated rules, set as classifiers, classify the process data. It is possible to use other techniques such as rules generalized from input data. In this case, we used random generated rules for simplicity. MDL criteria and the improvement rate evaluate these rules and classification results, and set the result to $CF_i$ with MDL or the improvement rate of MDL. Therefore, in the case of MDL criteria, each classifier is selected by MDL value as a maximization problem, and in the case of improvement MDL criteria, it is a minimization problem using the improvement rate of MDL. To each classifier, a new classifier is generated, conducting tournament selection based on the obtained $CF_i$, crossing and mutation.

When $\mathbf{p}$ is set as a conditional expression of the former part and $\mathbf{r}$ is set as a conditional expression of the latter part, the structure of the classifier becomes

$$\begin{aligned}\mathbf{p} &= (p_1^1 \wedge p_2^1 ... \wedge p_k^1) \vee (p_1^2 \wedge p_2^2 ... \wedge p_k^2) ...,\\ \mathbf{r} &= r_1, r_2, ..., r_n.\end{aligned}$$

The result $r_i$ shows all the possible results that the target event would obtain, and counts the number of hits in all $r_i$ that hit in the former part. This gives an estimated value of reliability for the latter part event, in accord with the agreed-upon event of the former part. The pressure toward generalization is preformed by the MDL metric.

### 3.4 Real Valued Learning Classifier System Using UNDX

Condition parts of the former section were four character values as 0-25, 25-50, 50-75, 75-100. The classifier system is extracted to real valued inputs in this section. The difference arises from changing the classifier condition. It is changed from a string:{0,1,2,,,#} to a concatenation of interval predicates:{(lower limit, upper limit),,,}, where each limit is a real value. A classifier matches an input $x$ if and only if $\mathit{lowerlimit} \leq x_i < \mathit{upperlimit}$, for all $x_i$.

This real valued LCS adopts an architecture of XCS, which is as follows: XCS is a recently developed learning classifier system that differs in several ways from

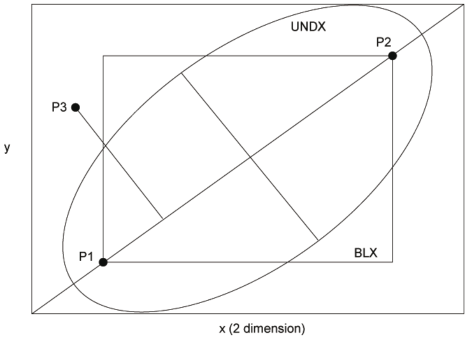


**Fig. 1.** UNDX

more traditional LCSs. In XCS, classifier fitness is based on the accuracy of a classifier's payoff prediction instead of the prediction itself. Second, the genetic algorithm takes place in the action sets instead of the population as a whole. Finally, unlike the traditional LCS, XCS has no message list and so is only suitable for learning in Markov environments. Although some real valued XCS have been proposed recently[11,12,13], those XCSs have been pointed out that the dependency between some data degrades the performance of the XCSs. So, this paper proposes a new method of real valued LCS based on UNDX[14]. UNDX means Unimodal Normal Distribution Crossover, which is shown at Figure 1. As shown at the figure, UNDX predicates new descendants with three parents. Additionally, this XCS based UNDX uses the method of MGG, which is one of generation change models of GA. MGG has high performance to maintain a diversity between genes. Our real valued LCS predicates continuous valued of conditions using the combination of UNDX and MGG.

The algorithm of UNDX is followings.

$$\begin{aligned}
&\mathbf{c_1} = \mathbf{m} + z_1\mathbf{e_1} + \sum_{k=2}^{n_{param}} z_k\mathbf{e_k}, \mathbf{c_2} = \mathbf{m} - z_1\mathbf{e_1} - \sum_{k=2}^{n_{param}} z_k\mathbf{e_k}, \\
&\mathbf{m} = (\mathbf{p_1} + \mathbf{p_2})/2, z_1 \sim N(0, \sigma_1^2), z_k \sim N(0, \sigma_2^2), \\
&\sigma_1 = \alpha d_1, \sigma_2 = \beta d_2\sqrt{n_{param}}, \\
&\mathbf{e_1} = (\mathbf{p_1} - \mathbf{p_2})/|\mathbf{p_1} - \mathbf{p_2}|, \mathbf{e_i} \perp \mathbf{e_j}(i, j = 1, \ldots, n_{param}, i \neq j).
\end{aligned}$$

where $n_{param}$ is the number of dimension, $\mathbf{p_1}, \mathbf{p_2}$ are parents, $\mathbf{c_1}, \mathbf{c_2}$ are kids, $d_1$ is the distance between the first and the second parents, $d_2$ is the distance between the third parent and the axis between the first and the second parents, $e_1$ is the unit vector of the axis between parents, $z_1, z_2$ are normal distribution, $\alpha, \beta$ are constant numbers which are $\alpha = 0.5, \beta = 0.35$, $e_k$ is a unit vector which is vertical to $e_1$.

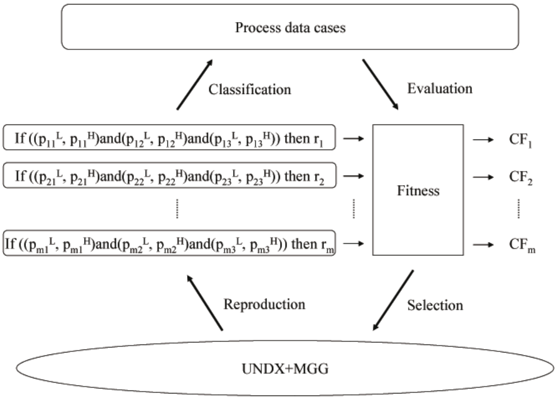

Process data cases
Classification
Evaluation
If (($p_{11}^L$, $p_{11}^H$)and($p_{12}^L$, $p_{12}^H$)and($p_{13}^L$, $p_{13}^H$)) then $r_1$
If (($p_{21}^L$, $p_{21}^H$)and($p_{22}^L$, $p_{22}^H$)and($p_{23}^L$, $p_{23}^H$)) then $r_2$
If (($p_{m1}^L$, $p_{m1}^H$)and($p_{m2}^L$, $p_{m2}^H$)and($p_{m3}^L$, $p_{m3}^H$)) then $r_m$
Fitness
$CF_1$
$CF_2$
$CF_m$
Reproduction
Selection
UNDX+MGG

**Fig. 2.** Learning classifier system

# 4 Experiments

## 4.1 Heuristic Search for Operation Rules

In the actual operation, it is significant to discover a control point that makes the final quality stable. The LCS with MDL criteria and the improvement rate, searches for the control rule targeting tag data with high correlation obtained by the process response model. Figure 3 shows an example of the classifier obtained. At this moment, the improvement rate is 3.1, and the MDL value is 32.9 bit. The next example shows the classifier in the case that considers the improvement rate. In this case, the improvement rate is 6.6, and the MDL value is 54.8 bit. The former becomes a simpler model, although the result is close to "commonplace" with a low improvement rate. In the case that gives consideration to the improvement rate, in addition to the MDL value, an unpredictable rule is more easily revealed. As a result of an interview with the person in charge of the operation, T2(temperature of the tower) provides an important control point that greatly affects product constituent quality in this biochemical plant, and it is too difficult to control the temperature. In order to control it more accurately, the classification rule that operators find hard to be aware of becomes precious information. In the case of Figure 3, F2 flow is expected to be related to T2 temperature, but it is not noticeable that F3 flow transition rate away from T2, or F4 flow are also connected to T2. Furthermore, in Figure 4.1 too, comparing

MDL+Improvement:
($25\% < F3 \leq 50\%$) and ($75\% < F4$) and ($F3$ is down) then $75\% < T2$
Real LCS:
($0.62 < F3 < 0.73$) and ($F4 < 0.55$) and ($F3$ is unstable) then $25\% < T2 < 50\%$
($0.28 < F1$) and ($0.57 < F2 < 0.69$) and ($F3 < 0.39$) and ($0.16 < F4$) and ($F1$ is up) and ($F3$ is up) and ($F4$ is down) then $75\% < T2$

**Fig. 3.** Control rules for Recipe

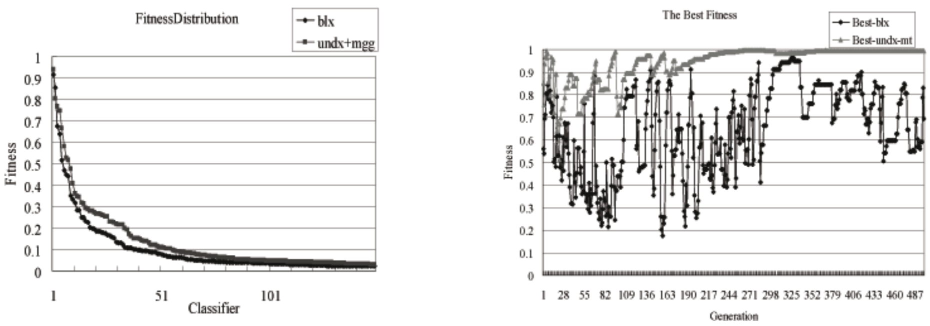


**Fig. 4.** Distribution of fitness(Left) and The Best fitness(Right)

with the F1 ingredient flow which directly infused into the distillation tower and the F2 return flow, operators found it hard to see that the F3 flow away from T2 is related to T2 temperature. As mentioned above, the application of Minimum Description Length criteria considering the improvement rate results in such unpredictable information.

### 4.2 Real Valued LCS Using UNDX

Figure 4 shows the distribution of fitness values of top 500 rules and the Best fitness values of each generation. Parameters of the experiment are population size = 4000, learning rate = 0.2, number of leaning steps = 40,000, number of descendants of MGG = 100.The XCS with UNDX and MGG reproduced more rules of that have better fitness than BLX in which descendants are generated by linear crossover.

## 5 Conclusion

This paper has proposed a new method of extracting plant operation knowledge from time series data using LCSs with the MDL principle and Tabu search. The method has generated useful but simple operation knowledge with high reliability. It enables us to extract implicit plant operation knowledge from both manual operation data and process data. Such knowledge is useful in transferring experts' special skills to inexperienced operators.

The method consists of the following phases: (1) Process data acquisition phase; (2) Normalization phase; (3) Reciprocal correlation analysis phase; (4) Process response model phase; (5) Extraction of a control rule phase; (6) Extraction of a workflow phase; (7) Variation analysis and improvement phase; and (8) Plant operation phase. Additionally, we have shown that extracted control rules of real-valued data.

The main contribution of the research is to establish a method of mining a set of meaningful control rules from the Learning Classifier System using the

Minimum Description Length criteria with improvement rate and real-valued LCS method. The effectiveness of the proposed method has been demonstrated using actual plant data. We believe that the proposed method is one of the practical LCS applications.

## References

1. Rissanen, J.: Modeling by shortest data description. Automatica 14, 465–471 (1978)
2. Adami, C.: Introduction to Artificial Life. Springer, Heidelberg (1998)
3. Mehta, M., Rissanen, J., Agrawal, R.: MDL-based decision tree pruning. In: Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD 1995), pp. 216–221 (1995)
4. Yamanishi, K.: A learning criterion for stochastic rules. Machine Learning 8, 165–203 (1992)
5. Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., Verkamo, A.I.: Fast discovery of assosiation rules. In: Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (eds.) Advances in Knowledge Discovery and Data Mining, pp. 307–328. AAAI Press and The MIT Press (1996)
6. Hilderman, R.J., Hamilton, H.J.: Knowledge Discovery and Measures of Interest. Kluwer Academic Publishers, Dordrecht (2001)
7. Holland, J.H., Reitman, J.S.: Cognitive systems based on adaptive algorithms. SIGART Bull. (63), 49 (1977)
8. Smith, S.: A learning system based on genetic adaptive algorithms. Ph.D thesis. University of Pittsburgh (1980)
9. Smith, S.: Flexible learning of problem solving heuristics through adaptive search. In: Proceedings 8th International Joint Conference on Artificial Intelligence (August 1983)
10. Butz, M.V., Pelikan, M., Llorà, X., Goldberg, D.E.: Extracted global structure makes local building block processing effective in XCS. In: GECCO 2005: Proceedings of the 2005 conference on Genetic and evolutionary computation, pp. 655–662. ACM, New York (2005)
11. Wilson, S.W.: Get real! XCS with continuous-valued inputs. In: Lanzi, P.L., Stolzmann, W., Wilson, S.W. (eds.) IWLCS 1999. LNCS (LNAI), vol. 1813, pp. 209–219. Springer, Heidelberg (2000)
12. Butz, M.V.: Kernel-based, ellipsoidal conditions in the real-valued XCS classifier system. In: GECCO 2005: Proceedings of the 2005 conference on Genetic and evolutionary computation, pp. 1835–1842. ACM, New York (2005)
13. Butz, M.V., Lanzi, P.L., Wilson, S.W.: Hyper-ellipsoidal conditions in XCS: rotation, linear approximation, and solution structure. In: GECCO 2006: Proceedings of the 8th annual conference on Genetic and evolutionary computation, pp. 1457–1464. ACM, New York (2006)
14. Ono, I., Kita, H., Kobayashi, S.: A real-coded genetic algorithm using the unimodal normal distribution crossover, pp. 213–237 (2003)

# Author Index